PROBABILISTIC AND GENERATIVE MODELING FOR EMOTION RECOGNITION AND SYNTHESIS

A social signal processing approach to socially intelligent agents in purposive group interactions

Navin Laxminarayanan Raj Prabhu

DISSERTATION

Probabilistic and Generative Modeling for Emotion Recognition and Synthesis

A social signal processing approach to socially intelligent agents in purposive group interactions

Kumulative Dissertation zur Erlangung des akademischen Grades
Dr. rer. nat.
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik
Universität Hamburg

eingereicht von

Navin Laxminarayanan Raj Prabhu

Hamburg 2025

Cover design by Mathimetha Vikram | `www.behance.net/mathimesuresh`

Navin Laxminarayanan Raj Prabhu: *Probabilistic and Generative Modeling for Emotion Recognition and Synthesis*

GUTACHTER:

Prof. Dr.-Ing. Timo Gerkmann

Prof. Dr. rer. nat. Nale Lehmann-Willenbrock

Prof. Dr. Carlos Busso

VORSITZ DER PRÜFUNGSKOMMISSION:

Prof. Dr. Frank Steinicke

Prof. Dr. Stefan Wermter

TAG DER EINREICHUNG:

04.12.2025

TAG DER DISPUTATION:

12.02.2026

Zusammenfassung

Artificial Intelligence (AI)-Technologien haben in den letzten Jahren bemerkenswerte Fortschritte gemacht und Systeme ermöglicht, die die Produktivität, Kreativität und Entscheidungsfindung des Menschen in verschiedenen Bereichen verbessern. Trotz dieser Fortschritte weisen aktuelle AI-Systeme nach wie vor große Defizite in Bezug auf soziale Intelligenz auf: die Fähigkeit, soziale Signale so wahrzunehmen, zu interpretieren und auszudrücken, dass eine nahtlose Interaktion zwischen Mensch und Agent unterstützt wird. Diese Einschränkung ist besonders kritisch bei *zielgerichteten sozialen Interaktionen*, bei denen der Erfolg von gegenseitiger Abhängigkeit, gemeinsamen Zielen, koordiniertem Handeln und gegenseitigem Vertrauen abhängt. In solchen Situationen kann ein Mangel an sozialem Bewusstsein die Qualität und die Ergebnisse der Interaktion grundlegend beeinträchtigen. Szenarien wie Besprechungen, Vorstellungsgespräche, kollaborative Design-Workshops und Gesundheitsberatungen erfordern, dass Agenten soziale Signale genau interpretieren, sich an zwischenmenschliche Dynamiken anpassen und angemessen reagieren, um eine effektive Zusammenarbeit aufrechtzuerhalten. Damit AI-Systeme in diesen Kontexten sinnvoll integriert werden können, müssen sie die Fähigkeit unter Beweis stellen, sozial angemessenes Verhalten zu verstehen und zu generieren. Diese Dissertation stützt sich auf das Rahmenwerk von *Social Signal Processing (SSP)*, das zwei wesentliche Komponenten sozialer Intelligenz unterscheidet: die *Erkennung* sozialer Signale und deren *Synthese*. Unter den vielfältigen sozialen Signalen – darunter Persönlichkeit, Dominanz, Rapport und Regulierung – spielt der *Affekt* eine zentrale Rolle. Als bewusste Gefühle wie Freude – Unlust oder Energie – Müdigkeit erlebt, durchdringt der Affekt soziale Interaktionen und prägt die Ergebnisse auf individueller, Gruppen- und Organisationsebene. Dementsprechend steht der Affekt im Mittelpunkt dieser Arbeit, wobei die Forschungsbeiträge sowohl seine Erkennung als auch seine Synthese umfassen.

Der erste Arbeitsbereich konzentriert sich auf die Erkennung affektiver Ausdrücke. Eine zentrale Herausforderung ergibt sich aus der inhärent subjektiven und mehrdeutigen Natur von Affektbezeichnungen – Annotatoren sind sich häufig uneinig, und die Zusammenfassung ihrer Urteile zu einer einzigen „Grundwahrheit" verwirft bedeutungsvolle Variationen. Um diesem Problem zu begegnen, entwickeln wir Methoden, die die Unsicherheit von Bezeichnungen für die Emotionserkennung auf individueller Ebene explizit modellieren. Durch die Kombination von Bayesian neural networks (BNN) mit Label Distribution Learning (LDL) erfasst unser Ansatz sowohl die zentralen Tendenzen als auch die Variabilität in den Annotationen, was zu Verbesserungen bei der Vorhersagegenauigkeit, der kalibrierten Unsicherheit und der Robustheit gegenüber Meinungsverschiedenheiten der Annotatoren führt. Darüber hinaus untersuchen wir, wie Unsicherheit unter begrenzten Annotationsbedingungen effektiv modelliert werden kann. Durch den Ersatz der Gaußschen Annahmen durch eine Student-t-Verteilung verbinden wir die Annotationssparsität direkt mit der Unsicherheit, was zu einer höheren Vorhersagegenauigkeit, einer schnelleren Konvergenz und einer verbesserten korpusübergreifenden Robustheit führt. Zusammen untermauern diese Ergebnisse ein zentrales Ziel dieser Arbeit: das Überdenken der Affektmodellierung, indem Unsicherheit nicht als zu

beseitigendes Rauschen betrachtet wird, sondern als eine intrinsische Eigenschaft affektiver Daten, die wertvolle Einblicke in emotionale Mehrdeutigkeit liefert. Während die Affekterkennung auf individueller Ebene bereits umfassend erforscht wurde, ist die kollektive Ebene des Affekts – wie Affekte auf Gruppenebene entstehen und sich innerhalb von Gruppen entwickeln – noch wenig erforscht. Um die Affektmodellierung von der individuellen auf die Gruppenebene auszuweiten, führen wir ein auf psychologischer Theorie basierendes Annotationsprotokoll ein, um die dynamischen Schwankungen des Gruppenaffekts zu erfassen. Auf der Grundlage dieser Annotationen schlagen wir ein graphbasiertes multimodales Framework vor, das kollektive Affekte in großem Maßstab modelliert, sowohl Konvergenz als auch Divergenz zwischen Gruppenmitgliedern erfasst und systematische Muster der Gruppendynamik aufzeigt.

Der zweite Arbeitsbereich befasst sich mit der Synthese affektiver Ausdrücke, wobei ein besonderer Schwerpunkt auf der Sprache liegt. Bisherige Ansätze stützen sich weitgehend auf gespielte Korpora mit parallelen Daten, die jedoch nicht die Vielfalt und Spontaneität natürlicher emotionaler Sprache erfassen können. Wir untersuchen generative Modelle für die Konvertierung von Sprachgefühlen in realen Daten unter Verwendung eines Entflechtungs-Resynthese-Frameworks: Durch die Trennung von Sprecher-, lexikalischen und affektiven Einbettungen und deren Rekonstruktion mit einem modifizierten HiFi-GAN-Vocoder erreichen wir eine Konvertierung ohne parallele Korpora. Dies führt zu Verbesserungen in Bezug auf Natürlichkeit, Steuerbarkeit und Tonhöhenmodulation. Wir untersuchen auch diffusionsbasierte Decoder als Alternative zu vocoderbasierten Methoden und zeigen, dass sie das Spektrum der Emotionen erweitern, die Ausdruckskraft verbessern und seltene Zustände besser erfassen, wenn auch manchmal auf Kosten der wahrnehmbaren Natürlichkeit. Schließlich untersuchen wir die Rolle prosodischer Hinweise wie Rhythmus und Betonung und integrieren einen speziellen Dauerprädiktor, der die Sprechgeschwindigkeit mit der Erregung moduliert und die Gesamtqualität und Natürlichkeit verbessert.

Zusammen fördern diese sieben Forschungsfragen sowohl die Aspekte der Wahrnehmung als auch der Reaktion sozial intelligenter Systeme. Auf der Seite der Erkennung leistet diese Arbeit einen Beitrag zu probabilistischen Methoden zur Modellierung von Label-Unsicherheit, führt neuartige Protokolle zur Annotation auf Gruppenebene ein und entwickelt skalierbare, multimodale graphbasierte Frameworks. Auf der Seite der Synthese schlägt sie unüberwachte generative Modelle für die Emotionskonvertierung in der Praxis vor, untersucht Diffusionsmodelle für die kontrollierbare Affektsynthese und demonstriert die Bedeutung der Dauer-Modellierung. Über technische Verbesserungen hinaus konzipieren diese Beiträge die Affekterkennung und -synthese neu, indem sie die Grundwahrheit als Verteilung statt als Absolutum behandeln, den Gruppenaffekt als dynamisch und emergent und die affektive Sprache als von Natur aus vielfältig und ökologisch begründet.

Durch die Auseinandersetzung mit diesen Herausforderungen legt diese Arbeit den Grundstein für KI-Systeme, die Affekte in zielgerichteten sozialen Interaktionen sowohl wahrnehmen als auch angemessen ausdrücken können. Solche Systeme versprechen eine Verbesserung des Vertrauens, der Zusammenarbeit und der Effektivität in verschiedenen gesellschaftlichen Bereichen, darunter Bildung, Gesundheitswesen und organisatorische Teamarbeit.

Abstract

Artificial Intelligence (AI) technologies have made remarkable progress in recent years, enabling systems that enhance human productivity, creativity, and decision-making across diverse domains. Despite these advances, current AI systems remain largely deficient in *social intelligence*: the ability to perceive, interpret, and express social signals in ways that support seamless human–agent interaction. This limitation is particularly critical in *purposive social interactions*, where success depends on interdependence, shared goals, coordinated action, and mutual trust. In such settings, a lack of social awareness can fundamentally compromise interaction quality and outcomes. Scenarios such as meetings, job interviews, collaborative design workshops, and healthcare consultations demand that agents accurately interpret social cues, adapt to interpersonal dynamics, and respond appropriately to sustain effective cooperation. For AI systems to be meaningfully integrated in these contexts, they must demonstrate the ability to understand and generate socially appropriate behavior. This dissertation adopts the framework of *social signal processing (SSP)*, which distinguishes two essential components of social intelligence: *recognition* of social signals and their *synthesis*. Among the wide array of social signals—including personality, dominance, rapport, and regulation—*affect* plays a central role. Experienced as conscious feelings such as pleasure–displeasure or energy–tiredness, affect permeates social interactions and shapes outcomes at individual, group, and organizational levels. Accordingly, affect constitutes the focus of this thesis, with research contributions spanning both its recognition and synthesis.

The first line of work focuses on the recognition of affective expressions. A central challenge arises from the inherently subjective and ambiguous nature of affect labels—annotators frequently disagree, and collapsing their judgments into a single "ground truth" discards meaningful variation. To address this, we develop methods that explicitly model label uncertainty for individual-level emotion recognition. By combining Bayesian neural network (BNN) with label distribution learning (LDL), our approach captures both the central tendencies and the variability in annotations, yielding improvements in predictive accuracy, calibrated uncertainty, and robustness to annotator disagreement. Furthermore, we investigate how uncertainty can be modeled effectively under limited annotation conditions. By replacing Gaussian assumptions with a Student's t-distribution, we directly connect annotation sparsity to uncertainty, resulting in higher predictive accuracy, faster convergence, and improved cross-corpus robustness. Together, these findings reinforce a central aim of this thesis: rethinking affect modeling by treating uncertainty not as noise to be eliminated, but as an intrinsic property of affective data that provides valuable insight into emotional ambiguity. While individual-level affect recognition has been extensively explored, the collective level of affect—how group-level affect emerges and evolves within groups—remains under-researched. To extend affect modeling from individuals to group-level, we introduce an annotation protocol grounded in psychological theory to capture the dynamic ebb and flow of group affect. Building upon these annotations, we propose a graph-based multimodal framework that models collective affect at scale, capturing both convergence and divergence across group members

and revealing systematic patterns of group dynamics.

The second line of work addresses the synthesis of affective expressions, with a particular emphasis on speech. Prior approaches largely rely on acted corpora with parallel data, which fail to capture the richness and spontaneity of natural emotional speech. We investigate generative modeling for speech emotion conversion in in-the-wild data using a disentanglement–resynthesis framework: by separating speaker, lexical, and affective embeddings and reconstructing them with a modified high-fidelity generative adversarial network (HiFiGAN) vocoder, we achieve conversion without parallel corpora. This yields improvements in naturalness, controllability, and pitch modulation. We also explore diffusion-based decoders as an alternative to vocoder-based methods, showing that they broaden the range of emotions, improve expressivity, and better capture rare states, albeit sometimes at the expense of perceptual naturalness. Finally, we study the role of prosodic cues such as rhythm and stress, integrating a dedicated duration predictor that modulates speech rate with arousal and enhances overall quality and naturalness.

Together, we frame seven research questions advance both the *perception* and *response* aspects of socially intelligent systems. On the recognition side, this thesis contributes probabilistic methods for modeling label uncertainty, introduces novel group-level annotation protocols, and develops scalable, multimodal graph-based frameworks. On the synthesis side, it proposes unsupervised generative models for in-the-wild emotion conversion, explores diffusion models for controllable affect synthesis, and demonstrates the importance of duration modeling. Beyond technical improvements, these contributions reconceptualize affect recognition and synthesis by treating ground truth as distributions rather than absolutes, group affect as dynamic and emergent, and affective speech as inherently diverse and ecologically grounded.

By addressing these challenges, this thesis lays the groundwork for AI systems that can both perceive and express affect appropriately in purposive social interactions. Such systems hold promise for enhancing trust, collaboration, and effectiveness across societal domains, including education, healthcare, and organizational teamwork.

Acknowledgements

Once, a wise person told me that achieving something meaningful requires three elements: *intention*, *growth*, and *blessing* (or destiny—whichever resonates individually). I wholeheartedly agree: one sows a seed (*intention*), nurtures it to help it flourish (*growth*), and, as this unfolds, hopes that unforeseen obstacles do not impede progress (*blessing*). I am deeply grateful to everyone who has contributed to making this dissertation a reflection of these three elements.

First and foremost, I owe my deepest gratitude to my supervisors, Timo Gerkmann and Nale Lehmann-Willenbrock. Thank you for being outstanding research advisors and for fostering the element of *growth* in this dissertation. Timo and Nale, I know that supervising an interdisciplinary thesis comes with its own challenges, but I cannot thank you both enough for giving me the freedom to explore and for trusting me to express myself through our research work. Having such kind, sincere, and approachable mentors is truly a privilege, and I am forever grateful for the past five years under your guidance, as well as for the personal relationship that has grown alongside it.

Timo, by leading through example, you have created an inspiring research environment of high standards and integrity in the SP Lab. Our discussions have continually pushed me to rediscover what defines 'good' research. Throughout this journey, your calm and steady guidance has helped me stay grounded during the highs and find direction and resilience during the lows.

Nale, your dedication to pushing the frontiers of interdisciplinary research has been instrumental in bringing this thesis project to fruition. I have always valued the trust you placed in me over the past five years—whether in the research directions I proposed or in my plans to hire additional annotators for data collection. You have consistently looked out for my development and career as a researcher. Our discussions have continually inspired me to grow and improve on multiple fronts. Thank you for everything.

My SP Lab colleagues—Alina, Danilo, Guillaume, Huajian, Jakob, Jean-Marie, Julius, Kristina, Lennart, Long, Martin, Rosti, Simon, Sina, and Tal: Thank you for creating such an inspiring and collaborative research environment. I have learned something unique from each of you, and I can confidently say that our lab exemplifies the highest standards of teamwork and intellectual engagement. Your contributions, insights, and brilliance in team discussions and peer reviews have continually sharpened this dissertation and enriched my research experience. Stephanie, thank you for your invaluable help with administrative matters and for always being there with your kind and caring conversations.

Vanessa, Masha, Marvin, and Clara: I am deeply grateful to you for being such inspiring interdisciplinary collaborators. Despite differences in disciplines, terminology, and research goals, you have been exceptional discussion and research partners, constantly challenging me to think beyond my field and encouraging me to venture outside my comfort zone. Working

with you has truly enriched my research perspectives. I hope our paths cross again in the future for more such research.

My family has been the solid rock supporting me throughout this journey. Their selfless sacrifices have endowed this dissertation with the element of *blessing*. To Amma and Daddy, I will be forever grateful and indebted to be your son. You have protected me with your prayers and provided me with the best, granting me the privilege of uninterrupted growth. My brother, Pranave, our conversations have always energized me and helped me rediscover myself. Amma, Daddy, and Pranave, your unwavering support, trust, and encouragement have been invaluable at every step of this journey.

Not to forget, I thank myself five years ago for upholding the *intention* to pursue this interdisciplinary research journey. Though nothing compares to the contributions of others in instilling *growth* and *blessing* into this dissertation, the unwavering *intention* was essential. Well done, Navin!

My beloved wife, Shri, the *wise person* in my life. You have seen it all and been through it all—you are my everything. Words cannot fully capture what you mean to me, nor the immense role you have played in this dissertation. This thesis is as much yours as it is mine, Kuttoos.

Forever grateful for the journey—the learning, the unlearning, and, most importantly, the friendships I have gained over the past five years. Thank you all again!

Table of Contents

Zusammenfassung i

Abstract iii

Acknowledgements v

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Purposive Group Interactions . . . 6
- 1.3 Affective Expressions in Group Interactions . . . 7
  - 1.3.1 Definition of Affect . . . 7
  - 1.3.2 Quantifying Affect . . . 8
  - 1.3.3 Conceptualization and Scope of Affect in Groups . . . 9
- 1.4 Social Signal Processing . . . 11
  - 1.4.1 The Modality of Speech . . . 13
  - 1.4.2 Social Signal Processing of Affect: Problem Formulation . . . 14
  - 1.4.3 Affect Datasets . . . 15
- 1.5 Recognition of Affect . . . 18
  - 1.5.1 Ground-truth of Affect [P1] . . . 20
  - 1.5.2 Probabilistic Modeling . . . 23
  - 1.5.3 Affect at the Group-level [P4] . . . 28
  - 1.5.4 Graph Modeling . . . 32
- 1.6 Synthesis of Affect . . . 34
  - 1.6.1 In-the-wild and non-parallel speech data . . . 36
  - 1.6.2 Generative Modeling [P9]–[P11] . . . 38
- 1.7 Outline and Contributions . . . 46

2 Overview of the Related Publications 49

3 Recognition of Emotional Expressions 52
- 3.1 Individual-level Emotion: Addressing Label Uncertainty Modeling . . . 53
- 3.2 Group-level Affect: Annotations and Multimodal Modeling . . . 73

4 Synthesis of Emotional Expressions 98
- 4.1 Speech Emotion Conversion using Neural Vocoder . . . 99
- 4.2 Speech Emotion Conversion using Diffusion Models . . . 105
- 4.3 Duration Modeling for Speech Emotion Conversion . . . 111

5 Discussion and Conclusions 119
- 5.1 Main Contributions of this Thesis . . . 119

5.1.1 Recognition of Affective Expressions 119
5.1.2 Synthesis of Affective Expressions 124
5.2 Directions for Future Research 128
5.3 Reflecting on Interdisciplinary Research and Collaborations 132

References **137**

List of Acronyms **157**

Eidesstattliche Versicherung **159**

A Related Peer-Reviewed Publications **163**
A.1 Bayesian Neural Networks for Label Uncertainty Modeling 164
A.2 *t*-distributions for Label Uncertainty Modeling 171
A.3 Leveraging Semantic Information for Speech Emotion Recognition 181
A.4 Ground-truth of affect labels 187

B Related Abstract Presentations **233**
B.1 Small Talk, Synchrony, and Entitativity 234
B.2 Group Synchrony and Attitude 247

1

Introduction

1.1 Motivation

Although Artificial Intelligence (AI)-powered systems have improved our lives by boosting productivity and creativity across many human tasks [1]–[5], they remain socially ignorant and lack *Social Intelligence* [6], [7]. Social Intelligence is the ability to express and recognize social signals and social behavior, an ability that humans posses and help them collaborate and co-create with other humans seamlessly [6], [8]–[10]. For instance, Zhou et al. [7] observe that general large language models (LLMs) often struggle to demonstrate social commonsense reasoning and strategic communication when interacting with human groups.

Not all intelligent systems need to be equipped with social intelligence. However, human-human or human-agent interactions are always socially situated [6], [9] and any intelligent system aimed to be seamlessly integrated within the context of socially situated interactions will need social intelligence [6], [10]–[12]. For example, research has explored agents that support collaboration in team meetings [13], systems embedded in educational environments to assist teachers in training students [14], and tools that aid psychotherapy by tracking bio-markers of stress or engagement to support psychiatrists in monitoring patient progress [15]. In such applications, agents that can detect social behaviors such as agreement, distraction, or disagreement, and respond appropriately with social cues in a polite, unobtrusive, and persuasive way are likely to be viewed as more natural, seamless, effective, and reliable [7], [9], [10], [16]. Ultimately, the ability to engage in socially situated interaction is not just a desirable feature but a prerequisite for the acceptance and long-term integration of intelligent systems into human society.

While social intelligence is crucial for AI integration within socially situated contexts in general, their impact can be even more profound specifically in *purposive social interactions*. As opposed to casual, spontaneous social interactions, purposive social interactions involve interdependence between the interacting partners to achieve a *shared purpose* through collaboration and co-creation [17]. Examples of purposive groups include a group of software developers, a debate, or a group of health workers assembled to brainstorm solutions to a problem. Within this context, socially intelligent agents can play a central role in fostering coordination, trust, and effective collaboration among participants [18], [19]. Such purposive interactions take a large part of our day-to-day lives. In an organizational context, an average employee spends approximately six hours per week in meetings, while managers dedicate around 23 hours weekly, with some even spending up to 80% of their time in meetings [20]. With their

significant presence across organizations, purposive interactions have a substantial impact on and are highly relevant to the functioning of many vital societal institutions, including industrial organizations, educational bodies, healthcare facilities, and government agencies [21]–[23]. Research on the integration of socially intelligent systems into purposive groups and its potential impact has been of interest to both computer scientists [7], [24] and researchers from the field of social science [13], [18], [25].

Works in the research field of social signal processing (SSP) [6], [11] suggest that, in order to achieve social intelligence, systems should be able to: (i) *model* and *analyze* social signals expressed by interlocutors during interactions, and (ii) *synthesize* social signals, thus allowing systems to generate appropriate responses. This distinction between understanding and generating social signals is foundational: before a system can produce socially appropriate behavior, it must first be able to perceive and interpret the signals conveyed by humans. In this sense, the analysis of social signals serves as the necessary basis upon which synthesis can build, and this understanding-synthesis loop provides a useful structure for approaching the study of social intelligence.

Among the many forms of social signals (listed in Table 1.1), *Affect* plays a particularly central role. Affect is generally defined as "a neurophysiological state[1] that is consciously accessible as a simple, non-reflective feeling (such as pleasure–displeasure, tension–relaxation, energy–tiredness)" [27], [28]. Permeating every purposive group interaction, *Affect* is an important social signal that shapes key individual-level (e.g., individual employee satisfaction, performance and turnover) as well as organizational-level variables (e.g., organizational climate; see [29], for an overview). Affect is also a crucial phenomenon for understanding group functioning and outcomes [29]–[31]. Affect shapes important processes and states such as creativity, prosocial behavior, satisfaction, decision making, teamwork, and leadership. Given its pervasiveness and relevance for individuals, groups, and entire organizations, understanding affect has been a central focus in social psychology research for nearly a century [32] and, more recently, has become a key research interest to computer scientists, too [33]–[35].

Building upon the above noted developments in the research fields of social science and computer science, the overarching aim of this thesis is to contribute towards equipping next generation computing systems with social intelligence, thereby enabling them to collaborate effectively in purposive social interactions. To realize this objective, we utilize SSP as the primary framework for modeling, analyzing, and synthesizing social signals, with a particular focus on *affect* as the central social signal. Following the SSP approach, specifically two research directions were investigated as part of the thesis[2]: (1) *Recognition* of affective expressions at the individual- and group-level, and (2) *Synthesis* of affective expressions. Recognizing affect at both the individual and group level is essential because purposive interactions unfold through the interplay between personal emotional states and emergent collective dynamics. For instance, in a brainstorming meeting, detecting an individual's growing frustration allows an agent to provide targeted support, while recognizing a broader decline in group enthusiasm may prompt interventions that re-energize the team as a whole. Addressing both levels ensures that socially intelligent systems can respond appropriately to the needs of individuals without losing sight of the overall group process. Figure 1.1 presents a group interaction scenario

[1] A neurophysiological state is a condition of the nervous system (especially the brain) that reflects ongoing patterns of neural activity, chemical signaling, and bodily regulation at a given time [26]. More on the relationship between the nervous system and the expressed affect is presented in Section 1.3.1

[2] For simplicity and following [33], we use the term *model* to cover both *recognition* and *synthesis* of affective expressions, whereas in SSP terminology recognition is typically framed as *modeling and analysis*.

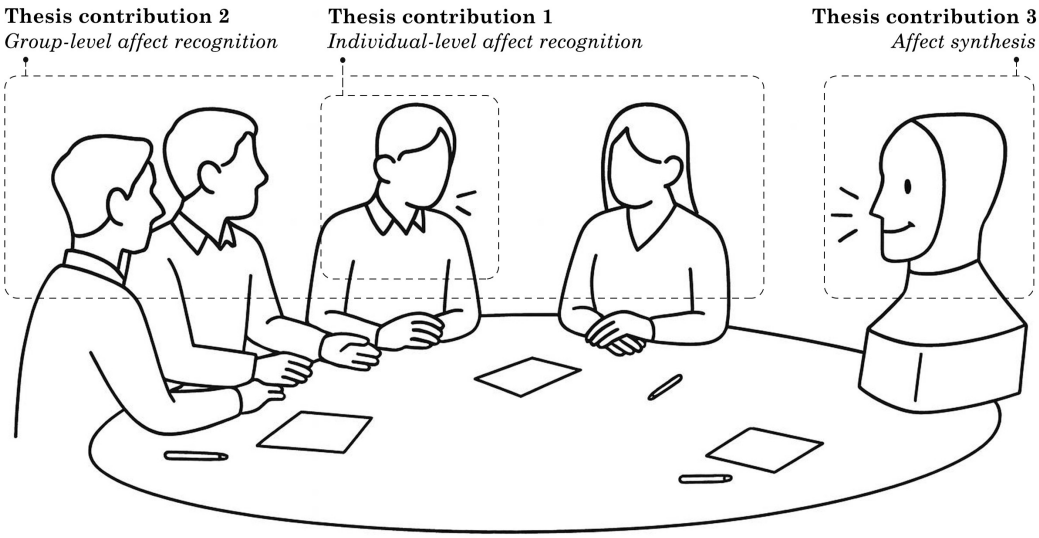


Figure 1.1: Example of a group interaction supported by a socially intelligent system illustrating the thesis contributions. **Thesis Contribution 1**: *Individual-level affect recognition* (Section 3.1), **Thesis Contribution 2**: *Group-level affect recognition* (Section 3.2), and **Thesis Contribution 3**: *Synthesis of affective expressions* (Chapter 4). The research questions are further detailed and presented in Section 1.7.

involving a socially intelligent system, which contextualizes the scope of this thesis.

Over the past 25 years, the emergence of data-driven modeling techniques, including machine learning (ML) and deep neural networks (DNNs), have acted as a catalyst to fuel rapid developments in the field of *Affective Computing* [33], [34], both in terms of affect recognition and its synthesis. The research field has seen a shift from traditional hand-crafted features and ML techniques based modeling to end-to-end DNN based approaches [35]–[38], with improved modeling capabilities in the respective tasks. However, a crucial challenge with research on affective computing in general, that draws interdisciplinary interest, is the availability of ground-truth affect labels that are both *ecologically valid* and *theoretically grounded* in social psychology literature [39], [40]. Such affect labels are essential for paired training data in data-driven techniques to accurately capture the naturalistic affective expressions exhibited by humans during real-world social interactions [41], [42]. These challenges with affect labels shape how we approach the two research directions of this thesis—affect recognition and affect synthesis.

***Recognition* of affective expressions** The first step towards social intelligence is the recognition of affective expressions, enabling systems to perceive and interpret affective expressions and processes displayed during an interaction. Under this task, treating affect as a multilevel social construct [43], we research on both individual-level and group-level affective expressions. *At the individual-level*, affect is a well researched topic [34], [35], [44], with well established datasets collected in natural, real world, and ecologically valid social settings. However, an often overlooked aspect in this context is the inherently subjective and ambiguous nature of affect labels [35], [45], [46]. Addressing this challenge is crucial for developing robust models that can adequately represent outliers or rare emotional categories, which are frequently dismissed, thereby improving generalization and fairness. This is because affective

responses are inherently variable across individuals and contexts, and failing to account for such variation can lead to biased predictions, misinterpretation of emotional states, and reduced reliability of the system in real-world interactions. To address this challenge, we adopt a *probabilistic* modeling approach [47], [48], moving beyond traditional deterministic methods. Probabilistic models are particularly well-suited for affective computing as they can capture the uncertainty inherent in affect labels—uncertainty that arises both from the subjective nature of human annotations [41], [46] and from the intrinsically ambiguous character of affect itself [45], [46]. A key advantage of probabilistic approaches over deterministic ones is their ability to generate *stochastic outputs* [47], [48]. The central motivation of this thesis is to leverage this capability, allowing affect labels to be represented as distributions of annotations rather than being simplified into a single averaged label, to better capture the variability, ambiguity, and richness of human affective responses, thereby improving model robustness, predictive accuracy, and fairness across diverse individuals and contexts.

While individual-level affect has been extensively studied, *group-level affect* remains comparatively underexplored [43], [49], [50]. In particular, modeling group-level affect in purposive social interactions represents a significant gap, even though organizational psychology has produced a substantial body of work on this topic over the past two decades [17], [30], [31], [51]. This gap has led to a lack of group-level affect annotations in the affect modeling literature, especially for purposive group interactions. To address this limitation, we collect group-affect labels in purposive interaction settings, using an annotation strategy and annotator training protocol grounded in organizational psychology research on group-level affect. Building on these labels, we then perform multimodal recognition of group-level affect.

By focusing on the automatic recognition of both individual- and group-level affective expressions, through addressing the associated challenges with respect to the ground truth labels of affect, our work contributes to the *perception* aspect of the broader goal of enabling socially intelligent agents within purposive social interactions.

***Synthesis* of affective expressions** Speech is one of the most fundamental social cue, not only in human-human interaction but also for interactions with socially intelligent agents [52], [53]. In this thesis, we particularly focus on speech-based synthesis of affective expressions, under the research domain of emotion-conditioned speech synthesis (ESS) [54], [55]. Similar challenges regarding the availability of ecologically valid data and reliable affect labels also arise in this research area. Much of the existing literature on speech-based affect synthesis relies on datasets of acted speech, which are often favored due to the relative simplicity and clarity of the performed affective expressions [56], [57]. Moreover such datasets have parallel speech samples, where each utterance of fixed semantic content is acted-out in different emotions by professional, trained actors [58]. Training affect synthesis models on such datasets fail to capture the spontaneous, diverse speaking styles of nonverbal cues like laughter and lip smacks, and disfluencies such as repetitions, hesitations, and interruptions [42], [57]. Empirical analyses using the NaturalVoices dataset [57] show that models trained on in-the-wild samples from MSP-Podcast [42] generate more natural and intelligible speech. Of note, moving away from acted-out speech samples necessitate the training strategy to not depend on parallel data samples.

We address the problem of *affective speech synthesis* for in-the-wild, real-world data using a *generative modeling* approach combined with an unsupervised training strategy, thereby avoiding reliance on parallel emotional speech samples. Advances in DNN-based generative methods—such as variational auto-encoders (VAEs) [59], generative adversarial networks

(GANs) [60], and the more recent diffusion models [61]—have driven significant progress in synthesis tasks across modalities, including image generation [62], speech synthesis [63], [64], and text generation [65]. However, in the context of unsupervised affective speech synthesis without parallel data, a key challenge persists: the *disentanglement* of speech attributes. Speaker identity, lexical content, and emotional content must be separated into mutually exclusive representations to enable effective conditional generation [55]. This thesis explores how generative modeling techniques, specifically *GANs* and *diffusion models*, can be adapted for affective speech synthesis in unsupervised settings, addressing both the lack of parallel data and the disentanglement problem in conditional generation.

By synthesizing individual-level affective expressions, and by addressing the associated challenges with respect to the *ecological validity* of the training data, our work contributes to the *response* aspect of the broader goal of enabling socially intelligent agents within purposive social interactions.

Organization of the Thesis The remainder of this introductory chapter outlines the scope and contributions of the thesis, with a focus on the two central research tasks: *affect recognition* and *affect synthesis*. It situates the work within existing literature and the broader development of the research landscape, which evolved in parallel with the contributions presented here. In doing so, it also introduces the core research questions addressed in the individual publications that comprise this cumulative thesis.

Section 1.2 begins with a review of foundational literature on *purposive social interaction* and its underlying dynamics. Section 1.3 then introduces *affect* as a crucial social signal, provides a formal definition of the construct, and motivates its importance for socially intelligent systems. Section 1.4 presents the research field of *social signal processing* (SSP), which offers a framework to study affect in a data-driven manner. This section also introduces key concepts related to *social signals and behavioral cues*, essential for understanding affect. Within this section, Subsection 1.4.1 highlights the role of *speech as a primary modality* for both recognizing and synthesizing affect, underlining its centrality in the development of socially intelligent agents. Subsequently, Subsection 1.4.2 formulates the specific task of affect modeling in the context of this thesis. Subsection 1.4.3 then provides a comprehensive overview of publicly available datasets for affect research, with particular attention to the nature of affect annotations. This forms the basis for subsequent discussions on the research tasks, which focus on challenges associated with affect labels and ground truth.

Next, Section 1.5 turns to the first research task of the thesis: *recognition of affect*. It emphasizes the use of *probabilistic modeling* for individual-level affect recognition and *graph-based approaches* for group-level affect recognition. Section 1.6 then shifts to the second task: *affect synthesis*. This section reviews relevant literature and introduces *generative modeling methods* that enable the production of emotionally expressive speech. Finally, Section 1.7 concludes the chapter by presenting the *overall structure of the thesis and corresponding research contributions*. It also describes how the included publications are organized and outlines their individual contributions.

The subsequent *chapters* present the core research contributions:

- Chapter 3 contains the accepted versions of articles focusing on *probabilistic modeling approaches* for affect recognition and analysis.
- Chapter 4 includes the accepted versions of articles centered on *generative modeling approaches* for affect synthesis.

Finally, Chapter 5 summarizes the key findings and contributions of the thesis (Section 5.1) and outlines promising *future research directions* (Section 5.2).

1.2 Purposive Group Interactions

A purposive group is "an intact social system, complete with boundaries, interdependence for some shared purpose, and differentiated member roles" [66]. Within such groups, it is possible to clearly distinguish members from non-members. More importantly, members rely on one another to achieve shared goals and, in the process, develop specialized roles and responsibilities within the group. In contrast, casual gatherings (e.g., people in a park or public square, spectators at a concert [67]) or spontaneous interactions (e.g., small talk in an elevator or waiting room, social events in a conference [68]) among individuals who lack a shared purpose do not constitute purposive groups. Purposive groups are prevalent in organizational settings, tasked with completing a wide variety of assignments across different time frames [17]. Examples of purposive groups include a team of software engineers collaborating on a product, or a group of healthcare professionals convened to develop strategies for addressing a medical challenge. These groups engage in structured and planned interactions, such as team meetings, collaborative problem-solving sessions, and coordinated task execution, all directed toward achieving a shared goal and purpose.

Group interactions[3] have long been a subject of interest for both social and organizational psychologists [22], [69]–[73]. This sustained attention stems from their ubiquity in organizational life and their central role in influencing how organizations function and perform [20]–[22], [29], [74]. Beyond organizational outcomes, group interactions are also closely tied to the well-being and satisfaction of individual members, shaping their motivation, stress levels, and overall sense of belonging [21], [51]. As discussed earlier, meetings constitute a substantial portion of employees' and managers' work time across organizational hierarchies [20], [75]–[77], underscoring the importance of understanding how these interactions unfold and impact group functioning. A growing body of research highlights the importance of studying group interactions—particularly in structured contexts such as team meetings—because they serve as key arenas for collective decision-making, coordination, knowledge sharing, and joint problem-solving [78]–[80]. Insights from multiple disciplines, including organizational psychology, communication studies, and computer-supported collaborative work, underscore that a deeper understanding of these interactional dynamics is essential for both theoretical advancement and practical improvement in organizational settings [81]–[83].

Earlier research on groups, largely based on self-reported measures, and predominantly outcome-oriented—aimed at identifying independent variables that statistically account for group-level outcomes [84]–[89]. However, such approaches often overlook the dynamic and temporal processes that underlie these outcomes [90]. In response, more recent research has adopted a process-oriented perspective, seeking to understand how events unfold over time, how causal mechanisms interact in complex or nonlinear ways, and how these dynamics contribute to emergent group outcomes [73], [91]–[93]. Such research has led social and organizational psychology researchers to code and analyze fine-grained, dynamic group processes such as leadership, humor, turn-taking, and conflict resolution, with increasing attention to how these micro-level behaviors unfold over time [40], [94]–[98]. Among these, *affect*—particularly how emotions are expressed, shared, and regulated in an interaction—is a key mechanism through which group processes and outcomes emerge [20], [29], [43], [99]–[101], and will be a central

[3]Throughout this thesis, we use the term "groups" to specifically represent purposive groups.

focus of investigation in this thesis.

1.3 Affective Expressions in Group Interactions

Group interactions are rich avenues for interpersonal expression and appraisal [31], [71], [102], where individuals continuously convey affect, intentions, and evaluations that dynamically influence how group members perceive and respond to one another in real time [102]. To examine such expressions and appraisals within group contexts, Vinciarelli et al. [6] propose a taxonomy that differentiates between *behavioural cues* and *social signals*. *Behavioural cues* refer to short-term temporal variations in neuromuscular or physiological activity—lasting from milliseconds to minutes—and include gestures, posture, facial expressions, prosody, and turn-taking, among others [6], [11]. *Social signals*, by contrast, represent expressions of an individual's social attitude, conveyed through combinations of these behavioural cues (e.g., affect, personality, dominance, cohesion, persuasion, etc.) [6], [11]. Further discussion of *behavioral cues* and *social signals*, including their distinctions and interrelationships, is provided in Section 1.4. These behavioural cues and social signals are both shaped by, and actively shape, the unfolding dynamics of group interaction. Notably, examining such dynamic behavioural cues and social signals in group interactions contributes to process-oriented research on purposive groups. As previously mentioned, this thesis concentrates on one particularly important dynamic social signal expressed during group interactions—*Affect*.

1.3.1 Definition of Affect

As an umbrella term encompassing both *feeling states* and *feeling traits* [29], affect is a pervasive phenomenon. Feeling traits capture an individual's predisposition to experience more positive or negative affect over time (i.e., positive or negative affectivity), whereas feeling states are dynamic and subject to short-term fluctuations. Feeling states include both diffuse moods and discrete emotions, which can be described along two well-established dimensions: *valence* (positive–negative) and *arousal* (level of activation) [29], [103], [104]. Moods are typically lower in intensity, longer in duration, and not directly linked to a specific stimulus [105], while emotions are generally more intense, short-lived, and elicited by specific events or situations [106]. In this thesis, we focus on emotional states and use the terms *emotion* and *affect* interchangeably[4].

Defining *emotion*, however, has been recognized as a notoriously difficult task [26]. Scherer [102], [107] addresses this challenge through a *component process* perspective, emphasizing that emotions involve multiple organismic subsystems acting in synchrony. Scherer's widely cited definition is as follows [26]:

> ***Emotion is an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems, elicited by the evaluation of an external or internal event as relevant to major concerns of the organism.***

In this framework, the five organismic subsystems comprise the central nervous system (CNS), the neuro-endocrine system (NES), the autonomic nervous system (ANS), the somatic nervous system (SNS), and the subjective feeling component. In simpler terms, this definition highlights that when a social event is appraised as personally relevant, it can trigger multiple,

[4]This interchangeable usage reflects a common trend in the literature, where individual-level affect is often referred to as *emotion*, whereas group-level affect is more frequently referred to as *affect*.

synchronized changes across different subsystems of the body and mind. For instance:

- The *CNS* is responsible for cognitive appraisal, attention, and conscious processing of the stimuli.
- The *NES* releases hormones that influence energy levels and stress responses.
- The *ANS* regulates physiological arousal, such as changes in heart rate or sweating.
- The *SNS* controls expressive behavior and bodily actions, such as facial expressions or posture.
- The *subjective feeling component* reflects the consciously experienced "feeling" of the emotion.

Together, these synchronized changes produce what we commonly recognize as an emotional episode. In this way, Scherer's framework highlights both the complexity and the multi-dimensional nature of emotions, while also explaining why they are difficult to define with a single criterion.

Research on affect as emotional states has highlighted three core characteristics: it is a (1) *dynamic*, (2) *multidimensional*, and (3) *multilevel* phenomenon [17], [29], [31]. First, emotional states are dynamic, meaning they are subject to short-term fluctuations in response to events and situational cues. Second, they are multidimensional, encompassing a broad range of both positive (e.g., happiness, joy, excitement) and negative (e.g., sadness, anger, fear, disgust) emotions. These emotions can be mapped along the dimensions of valence and arousal, as defined by the circumplex model [104] to which we will come to later in this section). Third, emotional states are multilevel phenomena: They can emerge and be observed not only at the individual level (e.g., a group member's individual emotion) but also at the group level (e.g., a shared emotional state within a group), particularly during group interactions.

Permeating every organizational level, *affect* shapes important individual-level (e.g., individual employee satisfaction, performance and turnover) as well as organizational-level variables (e.g., organizational climate; see [29], for an overview)). Affect is also a crucial phenomenon for understanding group functioning and outcomes. Indeed, group interactions are rich avenues of affective expressions and interpretation [29]–[31]. Affect shapes important individual and group processes such as creativity, prosocial behavior, satisfaction, decision making, teamwork, and leadership. Given its pervasiveness and relevance for individuals, groups, and entire organizations, understanding affect has been a central focus in organizational psychology research for nearly a century [32]. Over the past two decades, it has also attracted growing interest from computer scientists [33]–[35].

1.3.2 Quantifying Affect

The first step in studying affective expression is to quantify it, either through external annotations based on observers' perceptions or through self-reports from individuals involved in the group interaction. External annotations typically involve trained or naïve annotators observing the group interaction and labeling their perception of the affective expressions displayed by individuals [12], [40], [108]. A prerequisite for such perception-based annotations is that affect must be externally expressed through observable behavioural cues [6], [11]. However, a well-documented limitation of this approach is its inability to reliably capture the "true" experienced affect—particularly when such affect is not fully expressed [45]. This occurs, for example, when individuals surface-act to display socially desirable emotions [109], suppress

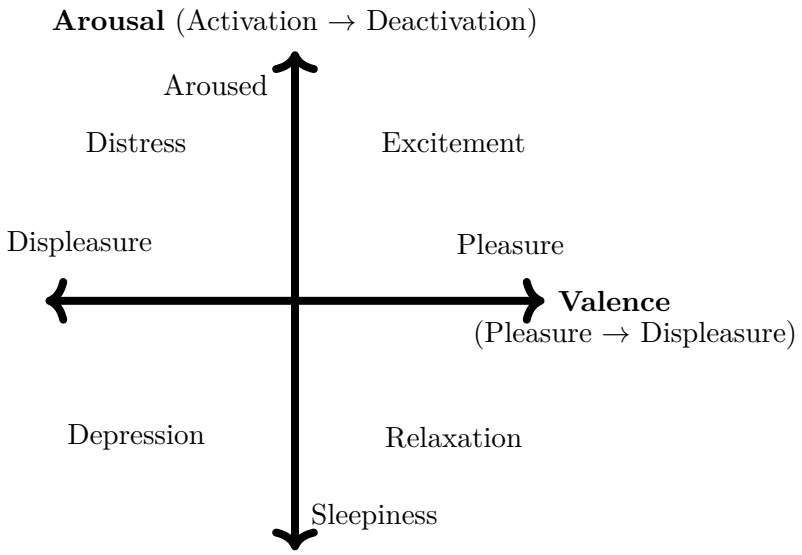


Figure 1.2: Circumplex Model of Affect illustrating Valence and Arousal dimensions.

their expressions in accordance with social norms [110], or regulate their emotional display for strategic purposes [111]. Despite this limitation, perception-based annotations remain widely used in the literature due to their resource-efficient setup and ease of deployment.

An alternative approach to quantifying affective experience involves using self-reports, wherein individuals report their internal affective states either during or after the interaction [112]. While this method is designed to access "true" experienced affect, it presents practical challenges. Post-interaction self-reports are susceptible to recall and recency biases [40], [108], while in-situ self-reports risk disrupting the natural flow and spontaneity of the interaction [40], [113]. To guide both human annotation and self-reporting of non-verbal affective expressions, researchers have traditionally relied on Ekman's six basic emotions theory [114], which defines six fundamental emotional categories: anger, disgust, fear, happiness, sadness, and surprise [106]. However, more recent research has highlighted the limitations of this discrete emotion framework, noting that affective expressions are often ambiguous and do not map cleanly to a single category [45], [104]. In response to these limitations, contemporary research has shifted toward continuous representations of affect, most notably the circumplex model [104].

The circumplex model represents affect within a two-dimensional space defined by *valence* (pleasure–displeasure) and *arousal* (activation–deactivation), conceptualized as orthogonal and bipolar dimensions. Figure 1.2 illustrates this model. By capturing the nuanced gradations and overlaps in affective experience, the circumplex model offers a more flexible and ecologically valid framework—particularly well-suited for studying affect in complex, dynamic social contexts such as group interactions [45], [91], [104].

1.3.3 Conceptualization and Scope of Affect in Groups

In this thesis, the construct of affect in groups is defined and approached in alignment with established theoretical perspectives, specifically with respect to its three core characteristics: (1) temporal dynamics, (2) multidimensionality, and (3) multilevel organization [17], [29], [31]. While prior works have addressed affect in various ways, several notable limitations remain. For example, many studies treat affect as a static construct (i.e., a feeling trait), without

modeling or examining its dynamic fluctuations over time [41], [42]. Similarly, affect is often conceptualized solely in categorical terms—focusing on discrete emotional states such as anger, happiness, or sadness—while overlooking its dimensional nature along continuous axes such as arousal and valence [58]. Finally, a more overlooked aspect is its *multilevel* nature: affect in groups is frequently examined only at the individual level, neglecting the emergent group-level affect that arises from patterns and dynamics of interaction between individuals [43].

Adopting the above conceptualization and scope of affect offers several important advantages and motivates treating it in this manner. First, modeling affect as a *dynamic* construct enables more robust representations that are better suited to real-world applications, as it can capture affective changes over time and adapt to evolving interaction contexts [43], [115]. Second, treating affect as a *multidimensional* construct along the arousal–valence domain provides a more flexible and comprehensive framework for representing affective expression [56], [104]. Such a perspective accommodates expressions that do not neatly fit into a single category and captures the gradations and overlapping characteristic of natural affective experiences—an especially valuable property when studying affect in complex social settings such as group interactions. Third, recognizing affect as a *multilevel* construct—particularly in the context of socially intelligent systems for purposive groups—supports a more holistic modeling approach, one that considers the well-documented links between affect and overall group functioning, success, and organizational outcomes [30].

Beyond these specific advantages, we argue that conceptualizing affect in this way is essential for three overarching reasons. First, it supports progress towards real-world capability and robustness, as the identified characteristics of affect reflect decades of empirical work aimed at accurately describing affective expression in natural settings [29], [104], [116]. Second, it achieves theory–method alignment by bridging theories of affect developed in social and organizational psychology with computational modeling methodologies from the fields of SSP and affective computing. Finally, it fosters truly interdisciplinary research, moving beyond the traditional *producer–consumer* model [117], in which social or organizational scientists produce data and annotations of observed team interactions, and computer scientists consume this information to develop modeling algorithms. Instead, we advocate for joint formulation of research questions and co-design of methodologies, thereby leveraging the insights and strengths of both disciplines to their fullest potential.

While conceptualizing affect in terms of its temporal dynamics, multidimensionality, and multilevel organization provides a theoretical foundation, studying affect in real-world group interactions requires operationalizing these constructs in observable behavior. The circumplex model emphasizes affect along continuous dimensions such as arousal and valence, but these internal states are not directly measurable; instead, they manifest through multiple social and behavioral channels—such as facial expressions, gestures, posture, speech prosody, and other paralinguistic cues. Capturing and interpreting these multimodal behavioral signals is therefore essential for understanding both individual- and group-level affect in social contexts. It is precisely this challenge that the interdisciplinary field of social signal processing (SSP) addresses: by enabling computational systems to sense, interpret, and generate social signals, SSP provides a framework for linking theoretical models of affect to observable behavior in group interactions.

1.4 Social Signal Processing

Social Signal Processing (SSP) is an interdisciplinary research field that aims to bridge the gap between human social intelligence and computational systems by enabling machines to sense, interpret, and generate social signals [6], [11]. Vinciarelli et al. [6] formally define *social signals* as communicative or informative cues that convey an individual's attitudes, emotions, intentions, and interpersonal relationships in social contexts. These signals are typically manifested and expressed through non-verbal *behavioural cues* such as facial expressions, gaze, posture, gestures, speech prosody, laughter, and other forms of paralinguistic communication. These social signals and behavioural cues are further illustrated in Figure 1.3. Subsequently, in Table 1.1 presents the social signals along with the behavioural cues that the psychologists consider the most important for conveying social information [118], [119].

Humans, in their day-to-day social interactions, are remarkably adept at understanding and managing the social signals expressed by those they communicate with [6]. This includes perceiving others' affective states and expressions, and responding with appropriate affective reactions and appraisals. This human capacity is referred to as *social intelligence* [6]. Social intelligence is considered a vital facet of human cognition—essential for building trust, maintaining cooperation, and navigating complex social environments. It underpins our ability to form relationships, collaborate effectively, and sustain social networks that are fundamental to both personal and collective well-being. The foundational works of Pentland [11], Vinciarelli et al. [6], and Picard [33] collectively contributed to the emergence of the field of *Social Signal Processing* (SSP), which aims to enable the next generation of computing systems with social intelligence. Notably, SSP provides a robust framework for studying affect in group interactions and paves the way for seamlessly integrating socially intelligent systems into such contexts. In this thesis, building upon arguments made in prior literature [6], [11], we argue that integrating socially intelligent systems into purposive group interactions can enhance their efficiency and overall effectiveness.

From an organizational and social psychology perspective, SSP of purposive group interactions enables fine-grained, process-oriented research on groups and their relationship with critical group outcomes and success [73], [93]. From a computer science perspective, SSP supports the development and integration of socially intelligent agents and systems into group interactions [6], [7]. Beyond discipline-specific motivations, SSP research opens avenues for impactful applications. For instance, one could envision a socially intelligent LLM participating in meetings as an active collaborator and co-creator. By equipping machines with the capability to process social signals, SSP facilitates the creation of socially-aware technologies such as emotionally intelligent virtual agents, socially adaptive robots, and affect-sensitive dialogue systems. These systems hold substantial potential for advancing human–computer interaction across domains including education, healthcare, customer service, and collaborative work environments [6], [7], [11], [120].

The central goal of SSP is to computationally model social signals, enabling machines to engage more naturally and effectively in human social interactions. This goal rests on two fundamental pillars [6], [11]: (1) the *recognition* of social signals, which supports accurate perception of social interactions and their underlying dynamics; and (2) the *synthesis* of social signals, which enables systems to respond with socially appropriate and contextually relevant behaviours, thereby facilitating effective participation. This work's focus on affect within an SSP framework also aligns with the broader domain of *Affective Computing*. The field, formally introduced by Picard [121], concerns the study and development of systems that

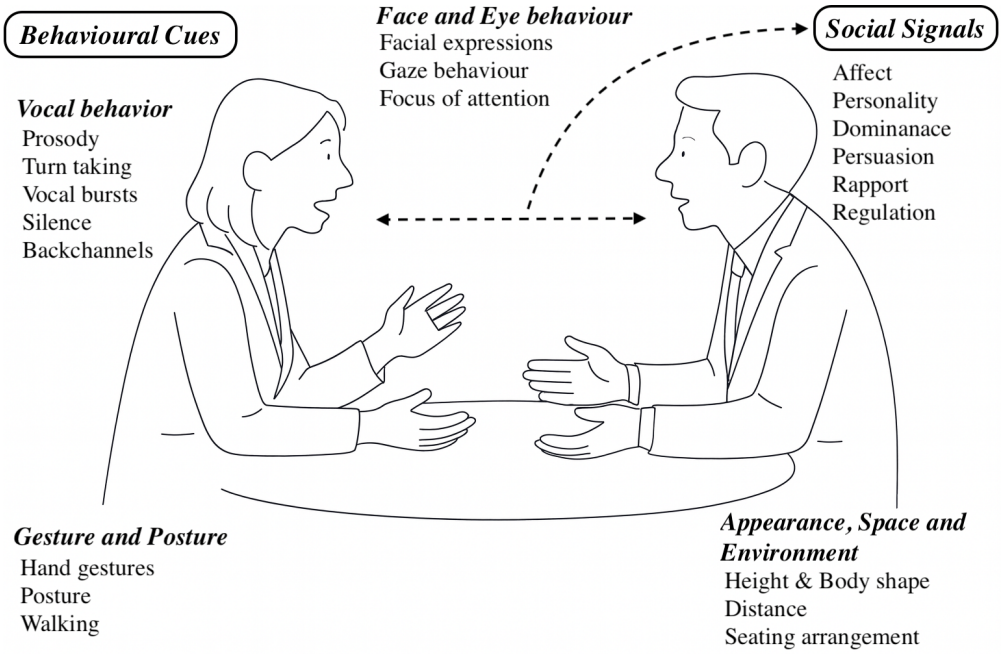


Figure 1.3: Taxonomy and Illustration of Behavioural Cues leading to Social Signals

recognize, interpret, and simulate human affective expressions. Affective Computing has since evolved into a widely studied interdisciplinary domain, encompassing emotion recognition, affect synthesis, and emotion-aware interaction in human–computer systems [122].

Approaches to SSP typically involve the automatic analysis and generation of both behavioural cues (e.g., gestures, prosody, posture) and the social meanings they convey (e.g., affect, dominance, engagement). These approaches span multiple modalities, including visual (e.g., facial expressions, gaze, body posture), acoustic (e.g., speech prosody, vocal bursts), and physiological (e.g., heart rate, skin conductance) signals, often in multimodal combinations to improve robustness and accuracy. As a multidisciplinary field, SSP integrates methods from computer vision, speech and audio processing, machine learning, psychology, and the social sciences to develop systems capable of interpreting and acting upon the social context of interactions.

Human affect, as a social signal, is expressed in group interactions through multiple modalities, both independently within each modality and jointly across modalities. These include: (i) *visual* cues in recorded video, such as gestures and posture [123], [124]; (ii) *acoustic* cues in recorded speech, such as prosody, linguistic content, and vocal bursts [35], [125]; and (iii) *physiological* cues, such as heart rate and skin conductance [126], [127]. While visual and acoustic recordings are relatively unobtrusive in group settings, the acquisition of physiological signals is often more intrusive [128]. Consequently, most SSP-based affect recognition studies rely on unimodal visual or acoustic methods, or on multimodal approaches that jointly leverage these two modalities [43]. Table 1.1 presents an overview of these multimodal cues along with their associated social signals.

In this thesis, we focus on speech as the key modality used for both modeling and synthesizing affect. While speech serves as the central modality, we also explore other communicative channels, including the *video* modality (i.e., Chapter 3.2), where visual cues complement

Table 1.1: Mapping between behavioural cues and social signals

	Social Signals				
Behavioural cues	**Affect**	**Personality**	**Dominance**	**Persuasion**	**Rapport**
Vocal behaviour					
Prosody	✓	✓	✓	✓	✓
Turn-taking	✓	✓	✓	✓	✓
Vocal bursts	✓	✗	✗	✗	✓
Silence	✓	✗	✗	✗	✓
Backchannels	✓	✓	✓	✓	✓
Face & Eye					
Facial expressions	✓	✓	✓	✓	✓
Gaze behaviour	✓	✓	✓	✓	✓
Focus of attention	✓	✓	✓	✓	✓
Gesture & Posture					
Hand gestures	✓	✓	✗	✓	✓
Posture	✓	✓	✓	✓	✓
Walking	✗	✓	✓	✗	✗
Appearance & Space					
Height	✗	✗	✓	✗	✗
Body shape	✗	✓	✓	✗	✗
Distance	✓	✓	✗	✓	✗

speech-based analysis by enriching the interpretation of social and affective dynamics. In addition, one of our studies (i.e., Section A.3) investigates the *text* modality, examining how semantic content contribute to understanding emotional expressions. The following sections discuss these modalities in greater detail and outline their relevance to SSP.

1.4.1 The Modality of Speech

Speech is the fundamental mode of human communication [129], supported by a unique combination of human anatomy and neural control [130]. Human speech is produced by sound generated in the larynx through vocal fold vibration and airflow, which is then filtered and shaped by the supralaryngeal vocal tract (SVT)—including the oral and nasal cavities—to create distinct and intelligible speech sounds [130]. The resulting formant frequencies—critical determinants of phonetic quality—enable the rapid and efficient transmission of complex linguistic and affective information. Affective states are conveyed not only through the choice of words but also through modulations in prosody, pitch, intensity, and timing—features that arise from dynamic control of the vocal folds in the larynx and the shaping of resonance in the vocal tract [116], [131]. This highly integrated system of articulatory gestures, prosodic variation, and acoustic encoding makes human speech an unparalleled vehicle for fast, precise, and emotionally expressive vocal communication. It enables the rapid and efficient exchange of ideas, emotions, and social signals, making it indispensable for coordinating action, building relationships, and maintaining social cohesion in both dyadic and group interactions [132]. Human speech can be broadly described in terms of its *prosodic* and *acoustic* attributes.

Prosodic attributes refer to suprasegmental features[5] such as pitch, intonation, rhythm, stress, and speech rate, which structure spoken language across phoneme and word sequences. These features play a central role in conveying affective, attitudinal, and interpersonal cues [116], [131], shaped simultaneously by linguistic constraints and the speaker's physiological and emotional state. *Acoustic* attributes, by contrast, capture the measurable physical properties of the speech waveform, including formant frequencies, spectral energy distributions, amplitude, and voice quality (e.g., breathiness, harshness, nasality). These arise from the configuration and movement of articulators, vocal tract resonances, and vocal fold vibrations, encoding both linguistic content and paralinguistic markers such as identity, health, and affective arousal [125], [132].

Together, prosodic and acoustic attributes provide complementary layers of information that make speech a uniquely rich channel for transmitting not only linguistic meaning but also affective states and social signals [6], [34], [116]. Importantly, speech is almost always available in human interactions and serves as the most fundamental cue through which intentions, emotions, and social dynamics can be observed and analyzed. While speech has long been central to human communication, it is now becoming equally critical for human–computer interaction. Recent advances in speech dialogue system (SDS) and speech language model (SLM) [52], [53] have positioned speech not just as an input modality for LLMs, but also as a medium for generating natural, expressive, and affect-aware spoken responses [133]–[136]. This dual capacity—to perceive intent, emotion, and context from speech, and to respond back in speech—brings systems closer to functioning as natural conversational partners. Yet, achieving robust social intelligence remains a major challenge: current systems must go beyond accurate linguistic parsing to reliably infer intent, recognize emotions, and produce contextually appropriate and socially engaging speech outputs [6], [7], [43], [53], [125], [135].

In this thesis, we take a SSP-oriented perspective to advance the development of socially intelligent systems, positioning speech as the primary channel of communication. An SSP-based approach, in this context, involves tackling both sides of the problem: the recognition and analysis of affect within group interactions, and the synthesis of affect-controllable speech to produce contextually appropriate, socially aligned responses. These two complementary dimensions are discussed in detail in Section 1.5 and Section 1.6, respectively.

1.4.2 Social Signal Processing of Affect: Problem Formulation

Formally, the task of social signal processing for affect can be defined as modeling the joint distribution between behavioral cues $\mathbf{X} \in \mathbb{R}^n$ and the corresponding social signal of affect $\mathbf{a} \in \mathbb{R}$:

$$P_{\mathbf{w}}(\mathbf{X}, \mathbf{a}), \tag{1.1}$$

where $\mathbf{w}$ denotes the parameters of the model. From this joint formulation, both affect recognition and affect synthesis naturally arise as conditional inference problems, i.e., $P_{\mathbf{w}}(\mathbf{a} \mid \mathbf{X})$ and $P_{\mathbf{w}}(\mathbf{X} \mid \mathbf{a})$, respectively. With respect to the three properties of affect described in Section 1.3.3—namely, (1) temporal dynamics, (2) multidimensionality, and (3) multilevel organization—the joint formulation requires further refinement:

Temporal Dynamics: Affect is inherently time-dependent. Accordingly, both the observed behavioral cues and the expressed affect should be modeled as temporal sequences within a

[5] In linguistics, *suprasegmental* features are properties of speech that extend over more than one sound segment (phoneme) and shape the overall contour of spoken language along time sequences.

context window:

$$P_{\mathbf{w}}(\mathbf{X}_{t_1:t_2}, \mathbf{a}_t)\,, \tag{1.2}$$

where $t_1 = t - w/2$ and $t_2 = t + w/2$ with w denoting the temporal context window and t the current time step.

Multidimensionality: Affect is expressed along multiple dimensions. Following Russell's circumplex model, affect is commonly represented in the arousal–valence space. Hence, the affect variable extends from a scalar to a vector:

$$\mathbf{a} = (a^{\text{arousal}}, a^{\text{valence}}), \quad \mathbf{a} \in \mathbb{R}^d, \tag{1.3}$$

where d is the dimensionality of the affect space. In the case of the circumplex model adopted in this thesis, $d = 2$.

Multilevel Organization: Affect can be modeled both at the *individual* and at the *group* level. Formally, the group-level affect modeling task can be expressed in terms of the joint distribution between the aggregated behavioral cues from all interactants and the corresponding group affect signal. Instead of explicitly enumerating each individual's cues, we define an aggregation function $h_\psi(\cdot)$, parameterized by ψ, which combines the multimodal behavioral signals of all group members into a joint representation. The formulation then becomes:

$$P_{\mathbf{w}}\Big(h_\psi(\{\mathbf{X}^{(i)}\}_{i=1}^{M}),\, \mathbf{a}^{(g)}\Big), \tag{1.4}$$

where $h_\psi(\{\mathbf{X}^{(i)}\}_{i=1}^{M})$ denotes the aggregated representation of behavioral cues from the M interactants, and $\mathbf{a}^{(g)}$ is the group-level affect.

Unified Formulation: In summary, by integrating the three key characteristics of affect, affect can be modeled as a joint distribution that accounts for the time-varying (Eq. (1.2)), multidimensional (Eq. (1.3)), and multilevel nature (Eq. 1.4) of affective expressions. Accordingly, the unified formulation of affect modeling is expressed as follows:

$$\text{Individual-level: } P_{\mathbf{w}}\Big(\mathbf{X}_{t_1:t_2}^{(i)},\, \mathbf{a}_t^{(i)}\Big)\,, \quad \mathbf{a}_t^{(i)} \in \mathbb{R}^d, \tag{1.5}$$

$$\text{Group-level: } P_{\mathbf{w}}\Big(h_\psi\Big(\{\mathbf{X}_{t_1:t_2}^{(i)}\}_{i=1}^{M}\Big),\, \mathbf{a}_t^{(g)}\Big), \quad \mathbf{a}_t^{(g)} \in \mathbb{R}^d, \tag{1.6}$$

where $\mathbf{X}_{t_1:t_2}$ and $\mathbf{a}_t$ denote the behavioral cues and the respective affect within a temporal context window, respectively, and d is the dimensionality of the affect space (e.g., arousal–valence). This formulation can thus be instantiated either at the *individual* level, where $\mathbf{a}^{(i)}$ denotes the affect of a single interactant, or at the *group* level, where $\mathbf{a}^{(g)}$ represents the collective affective state of the group.

1.4.3 Affect Datasets

The aforementioned characteristics of affect directly guided the selection of emotion datasets for the tasks of recognition and synthesis. An additional criterion considered in this process was the degree to which datasets capture *in-the-wild* conditions. In group settings, emotion datasets can broadly be categorized into two types depending on the nature of the interactions they record: (i) *acted-out datasets*, which rely on affective expressions elicited in scripted, staged interactions [41], [58], and (ii) *in-the-wild datasets*, which capture affective expressions in more naturalistic group interaction contexts [42], [115]. While acted-out datasets are

Table 1.2: Comparison of datasets for *individual*-level affect recognition

Dataset	Group Setting	Group Size	Temporal Dynamics	Affect Annotations	Modalities	Ecological Validity
AMI (2005) [137]	Meetings of design task	4	✗	Categorical affect	Audio, Video	Acted and In-the-wild
GAMEON (2020) [138]	Escape game task	3-4	✗	Categorical affect	Audio, Video, Motion Capture	In-the-wild
IEMOCAP (2008) [41]	Scripted roleplay and spontaneous actors	2	✗	Arousal-Valence	Audio, Video, Motion Capture	Acted
MSP-Improv (2017) [139]	Acted conversations	2	✗	Categorical affect	Audio, Video	Acted
MuSe-CaR (2017) [140]	Youtube monologue reviews	2	✓ $w = 100$ ms	Arousal-Valence	Audio, Video	In-the-wild
SEMAINE (2011) [141]	Human-Agent collaboration	2	✓ $w = 100$ ms	Arousal-Valence	Audio, Video	In-the-wild
AffWild2 (2020) [142]	Youtube monologue videos	1	✓ $w = 100$ ms	Arousal-Valence	Video, Audio	In-the-wild
RAVDESS (2018) [143]	Scripted speech and song	1	✗	Categorical affect	Audio, Video	Acted
CREMA-D (2014) [144]	Crowd-acted emotional speech	1	✗	Categorical affect	Audio, Video	Acted
CMU-MOSEI (2018) [145]	Online monologues	1	✓ $w = 100$ ms	Categorical affect	Audio, Video	In-the-wild
AMIGOS (2018) [146]	Short/long video watching	1–4	✓ $w = 100$ ms	Arousal-Valence, Personality	Video, Physiological	Controlled/In-the-wild
AffectNet (2017) [147]	Internet facial images	1	✓ $w = 100$ ms	Arousal-Valence, Expressions	Image	In-the-wild
MELD (2019) [148]	Dyadic/multiparty dialogues (Friends TV show)	2–5	✗	Categorical affect, Sentiment	Audio, Video	Acted
CAST-Phys (2025) [149]	Individual physiological stimuli	61	✗	Valence-Arousal	Video, Physiological	In-the-wild
EAV (2024) [150]	Dyadic cue-based conversation	2	✗	Categorical affect	Physiological, Audio, Video	Acted
Mixed-Emotion (2024) [151]	Individual video watching	1	✗	Categorical affect	Video, Physiological	Acted
EVA-MED (2025) [152]	Individual stress induction	1	✗	Arousal-Valence	Physiological	Acted
AVES (2025) [153]	Individual video watching	1	✓ $w = 100$ ms	Categorical affect	Audio, Video	In-the-wild
EmotionTalk (2025) [154]	Dyadic dialogue (Chinese)	2	✗	Categorical affect	Audio, Video	Acted
MSP-Conversation (2020) [115]	Podcast conversations	2-5	✓ $w = 60$ ms	Arousal-Valence	Audio	In-the-wild
SEWA (2019) [155]	Dyadic online conversations	2	✓ $w = 100$ ms	Arousal-Valence	Audio, Video	In-the-wild
MSP-Podcast (2018) [42]	Podcast conversations	2-5	✗	Arousal-Valence	Audio	In-the-wild
RECOLA (2013) [156]	Online collaborative task	2	✓ $w = 40$ ms	Arousal-Valence	Video, Audio	Spontaneous

valuable for controlled experimentation, models and inferences built solely on such data may struggle to generalize to the complex, diverse, and nuanced affective processes that characterize real-world group interactions [35], [42], [56].

A notable pitfall of acted-out datasets is their reliance on scripted scenarios and instructed performances, which often produce exaggerated or prototypical affective displays. As a result, they lack the spontaneity and contextual grounding of natural interactions, limiting the ecological validity of models trained on them. By contrast, *in-the-wild* datasets offer greater realism, capturing the subtleties and diversity of affective behaviors as they occur in authentic group dynamics [42]. This ecological validity makes them indispensable for building models that generalize beyond controlled conditions [157], [158]. Nevertheless, collecting in-the-wild data is inherently challenging due to noisy and uncontrolled environments, technical limitations in multimodal recording, and the considerable annotation effort required. Despite these obstacles, the advantages of in-the-wild datasets in enabling robust, generalizable, and contextually meaningful affect modeling outweigh their challenges [57].

Table 1.3: Comparison of datasets for *group*-level affect recognition

Dataset	**Group Setting**	**Group Size**	**Temporal Dynamics**	**Affect Annotations**	**Modalities**	**Ecological Validity**
HAPPEI (2012) [159], [160]	Social events (party, marriage, convocation)	4	✗	Happiness intensity (6 levels)	Images	In-the-wild
MultiEmoVA (2015) [161]	Social events (sports, crowds)	14	✗	Arousal–Valence (3 levels each)	Images	In-the-wild
DynamicAffect (2015) [91]	Team meetings	2	✓ $w = 2$mins	Arousal–Valence (continuous)	Video	Controlled
GAF 2.0 (2015) [49]	Social events (party, marriage, convocation)	3	✗	Affect (3 classes: pos/neg/neutral)	Images	In-the-wild
GAF 3.0 (2018) [162]	Social events (party, marriage, convocation)	3	✗	Valence (3 levels)	Images	In-the-wild
Group Cohesion (2019) [163]	Social events (wedding, family, birthday)	5	✗	Valence (3 levels)	Images	In-the-wild
AloneVsGroup (2019) [164]	Groups engaging with multimedia content	3	✓ $w = 20$s	Arousal–Valence (continuous)	Video	Controlled
VGAF (2020) [165]	YouTube videos (protest, festival, wedding, fighting)	3	✗	Valence (3 levels)	Video	In-the-wild
GroupEmoW (2020) [166]	Social events (party, protest, wedding)	5	✗	Valence (3 levels)	Images	In-the-wild
SiteGroEmo (2022) [167]	Travel destination groups	5	✗	Valence (3 levels)	Images	In-the-wild
GECV (2022) [168]	Crowd events (festival, parade, funeral)	3	✗	Valence (3 levels)	Video	In-the-wild
MIP-GAF (2025) [169]	Social gatherings (funeral, concert, wrestling, etc.)	2 + 1 LLM	✗	Most Important Person + Valence	Images	In-the-wild

Through Tables 1.2 and 1.3, we present a consolidated summary and tabular comparison of affect datasets frequently employed in the affective computing and emotion research literature. The emphasis is placed on datasets that provide explicit affect annotations and are obtained

within purposive group contexts. Table 1.2 enumerates datasets pertinent to *individual*-level affect modeling, whereas Table 1.3 outlines datasets specifically developed for *group*-level affect modeling.

Referring to Table 1.2, we find that only a limited set of datasets—namely *MSP-Conversation* [115], *SEWA* [155], and *RECOLA* [156]—satisfy our requirements of being collected in *in-the-wild* contexts involving purposive group interactions. Crucially, the affect annotations provided in these datasets align with the two defining characteristics of affect that form the focus of our study: they offer temporally continuous annotations that capture the *dynamic* nature of affect, and they include multidimensional descriptors (e.g., valence and arousal) that reflect its *multidimensional* nature. Among these, however, access to the *SEWA* dataset was not possible due to administrative constraints. Accordingly, throughout this thesis, we rely primarily on *MSP-Conversation* [115] and *RECOLA* [156] to investigate individual-level affect, as presented in Section 3.1. In addition, although the MSP-Podcast [42] dataset does not provide temporally continuous annotations and is instead restricted to utterance- and segment-level labels, it satisfies the other requirements and is therefore employed for the task of affect synthesis, as presented in Chapter 4. In this setting, temporally continuous annotations are not essential because temporal dynamics can be controlled through the independent synthesis of successive segments, allowing the desired affective progression to be specified explicitly rather than inferred from continuous labels.

In contrast, as shown in Table 1.3, none of the existing group-level affect datasets fulfill these criteria. While they serve as useful benchmarks for group-level affect recognition, they do not simultaneously provide *in-the-wild* purposive group interaction data together with temporally dynamic and multidimensional affect annotations. To address this gap, we collected a new dataset explicitly designed to enable the study of affect at the group level, as further discussed in Section 3.2.

1.5 Recognition of Affect

Affect *recognition* has been predominantly examined in the computer science literature under the task termed automatic emotion recognition (AER), irrespective of the modality employed. When limited to a single modality, it is instead categorized as facial emotion recognition (FER), speech emotion recognition (SER), or physiological emotion recognition (PER), corresponding to facial, speech, and physiological signals, respectively[6]. It is important to note that all of these tasks pertain exclusively to individual-level affect recognition. At the group level, the corresponding modeling task is referred to as group emotion recognition (GER), regardless of the modality utilized. This chapter introduces *affect recognition* as the foundational step toward equipping systems with *social intelligence*—the capability to interact more naturally with humans by *perceiving, recognizing, and interpreting* the emotional states of individuals within social contexts.

Affective and emotional states can be expressed by individuals within a group, either voluntarily or involuntarily, in manners that may be consciously controlled or entirely unconscious. Such expressions may be overtly visible—such as a smile—or perceptible only through close physical contact, as in the warmth of a reassuring handshake. Emotions can also be conveyed through actions, for instance, offering comforting words to someone in distress. In all these cases,

[6]In the computer science literature on affective computing—unlike in the psychology literature—the term "emotion" is more commonly preferred to denote affect recognition, which explains the use of the suffix "-ER" in these acronyms.

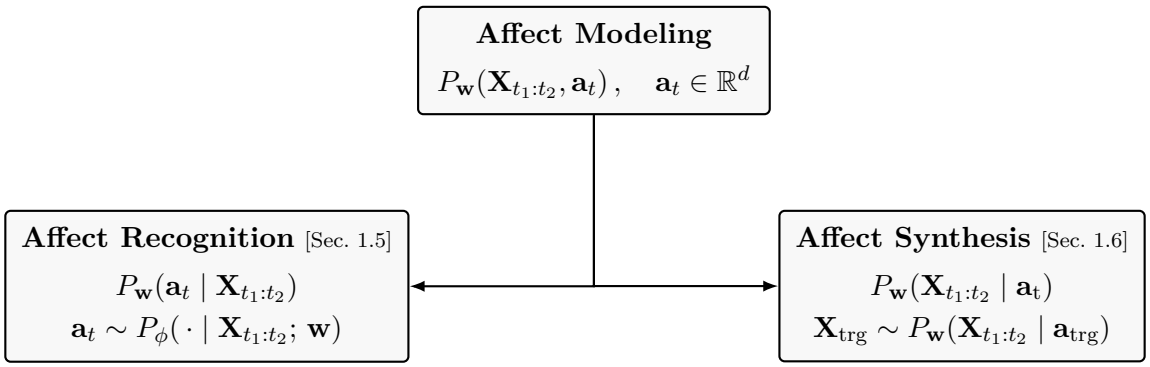


Figure 1.4: General form of affect modeling (top) categorized into two complementary tasks: affect recognition (bottom-left) and affect synthesis (bottom-right).

patterns of information are transmitted via behavioral cues across multiple modalities, and these patterns can be modeled and represented by computational systems. The concept of a "model" in this context can be understood as a function that learns a mapping—specifically, a mapping between behavioral cues and the underlying affective states they signify—by capturing momentary or recurring patterns in those cues.

As noted in Figure 1.4, with the modeling of affect, the task of *affect recognition* can be formulated as the conditional inference of the affective state given observed behavioral cues. This is expressed as

$$P_{\mathbf{w}}(\mathbf{a}_t \mid \mathbf{X}_{t_1:t_2}), \tag{1.7}$$

where the task is to estimate $\mathbf{a}$ from $\mathbf{X}$. In practice, this conditional distribution is approximated by a data-driven mapping $f_{\mathbf{w}}(\cdot)$ parameterized by w:

$$\mathbf{a}_t = f_{\mathbf{w}}(\mathbf{X}_{t_1:t_2})\,, \quad \hat{\mathbf{a}}_t \in \mathbb{R}^d, \tag{1.8}$$

where the parameters $\mathbf{w}$ are learned in a supervised manner by minimizing the prediction error between estimated and annotated affect values:

$$\mathbf{w}^\star = \arg\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_{\mathbf{w}}(\mathbf{X}_i),\, \mathbf{a}_i)\,, \tag{1.9}$$

where $\mathcal{L}(\cdot)$ denotes the emotion recognition specific loss function. Typically, in the task of affect recognition, the concordance correlation coefficient (CCC) [170] has been widely used as the loss function, $\mathcal{L}(\cdot)$, [34], [35]. The CCC metric evaluates the agreement between the ground-truth $\mathbf{a}$ and its estimate $\hat{\mathbf{a}}$. Its values range from -1 to $+1$, with $+1$ indicating perfect concordance. Unlike the Pearson's correlation coefficient (PCC), which measures only the strength of linear association, the CCC accounts for both correlation and mean bias, making it more suitable as both an evaluation and loss function in affect recognition and modeling. The PCC between $\mathbf{a}$ and $\hat{\mathbf{a}}$, over T frames, is given by:

$$r = \frac{\sum_{t=1}^{T}(\mathbf{a}_t - \mu_{\mathbf{a}})(\hat{\mathbf{a}}_t - \mu_{\hat{\mathbf{a}}})}{\sqrt{\sum_{t=1}^{T}(\mathbf{a}_t - \mu_{\mathbf{a}})^2}\sqrt{\sum_{t=1}^{T}(\hat{\mathbf{a}}_t - \mu_{\hat{\mathbf{a}}})^2}}, \tag{1.10}$$

where $\mu_{\mathbf{a}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{a}_t$ and $\mu_{\hat{\mathbf{a}}} = \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{a}}_t$ denote the means of the ground-truth and

predicted values, respectively. For a given PCC (r), the CCC between $\mathbf{a}$ and $\hat{\mathbf{a}}$ is then expressed as:

$$\mathcal{L}_{\text{CCC}}(\mathbf{a}, \hat{\mathbf{a}}) = \frac{2r\,\sigma_{\mathbf{a}}\sigma_{\hat{\mathbf{a}}}}{\sigma_{\mathbf{a}}^2 + \sigma_{\hat{\mathbf{a}}}^2 + (\mu_{\mathbf{a}} - \mu_{\hat{\mathbf{a}}})^2}, \tag{1.11}$$

where $\sigma_{\mathbf{a}}^2 = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{a}_t - \mu_{\mathbf{a}})^2$ and $\sigma_{\hat{\mathbf{a}}}^2$ is defined analogously for $\hat{\mathbf{a}}$.

Similarly, for the modeling function $f_{\mathbf{w}}(\cdot)$, two broad categories of approaches have been adopted in the literature: (i) traditional hand-crafted feature extraction methods [171], [172], and (ii) end-to-end learning frameworks [173]–[175]. In the feature extraction paradigm, the pipeline typically involves first extracting hand-crafted speech or visual descriptors, then fusing these features, and finally applying data-driven supervised models such as support vector machine (SVM) or multi-layer perceptron (MLP), as well as DNN-based architectures, to achieve the mapping defined in (1.8). More recently, however, there has been a significant shift towards leveraging self-supervised learning (SSL)-based representations obtained from large-scale pretraining, such as ResNet [176], VGGish [177], *wav2vec* [178], *HuBERT* [179], *WavLM* [180], and related models. These representations are either fine-tuned or employed in a frozen manner for downstream affect recognition tasks, offering a powerful alternative to traditional hand-crafted features and improving robustness in real-world scenarios [37], [38], [181], [182]. While this setup reflects the conventional approach in affect recognition [34], [35], [183], the field of AER has advanced substantially, with notable progress in the last two decades [182]. A key shift has been the move from scripted, acted datasets toward *in-the-wild* corpora that capture more naturalistic affective behavior [42], [115]. Parallel to this, research has advanced in modeling affect's defining properties: its temporal *dynamics* [43] and its *multidimensionality*, increasingly studied through the circumplex model (Figure 1.2) [37], [181]. Despite these developments, two challenges remain central to AER:

Two Major Challenges

1. Defining a reliable "*ground-truth*" for affect modeling (Section 1.5.1)
2. Accounting for the "*multilevel*" nature of affect (Section 1.5.3)

1.5.1 Ground-truth of Affect [P1]

In the tasks of AER, SER, FER, and PER, for the data-driven mapping (1.8), the establishment of dependable ground-truth labels is arguably of paramount importance [35]. "Ground-truth" designates the accurate and authoritative data or labels that provide a reference standard for training and evaluating such approaches. It denotes the correct outcome or label for each individual instance within a dataset [184], [185]. Within AER, ground-truth typically consists of non-verbal social cues signaling affect. These labels are frequently obtained from human annotation and form the cornerstone for modeling the relationship between behavioral cues and affect. Therefore, the precision of human annotations underpins the development of automatic affect recognition systems [34], [45].

A crucial challenge in defining the "ground-truth" originates from the intrinsically subjective and ambiguous nature of affect. *Subjectivity* stems from the deeply personal and interpretive nature of affect. Since affect is predominantly assessed through observable emotional expressions and quantified by external annotators, the resulting labels inevitably embed the annotator's individual interpretation [35], [46], [186]. In contrast, *ambiguity* emerges due to the fact that an affective stimulus can elicit several, simultaneously present, yet mutually exclusive

affective categories [45], [46]. Together, the subjectivity and ambiguity in labeling affective expressions drive a broader, inherent *uncertainty* in affect annotations. This uncertainty encapsulates both the lack of annotator confidence and the imprecision in assigning affective labels to stimuli [46]. Accordingly, throughout this section and the thesis, the term *uncertainty* serves as an overarching term encompassing both the *subjectivity* and *ambiguity* intrinsic to affect annotations.

To mitigate the impact of such uncertainty, annotations are typically obtained from multiple annotators, with the "ground-truth" subsequently derived through aggregation strategies such as averaging [42], majority voting [187], or evaluator-weighted estimation (evaluator-weighted means (EWE)) [188]. While these approaches aim to improve reliability, they may also discard valuable information about the inherently subjective nature of affective expression, thereby obscuring subtle and nuanced affective cues [35]. In practice, robust modeling of affect requires going beyond consensus-based "true" labels to explicitly account for the uncertainty present in human annotations [35]. Formally, given K annotators providing annotations $\{\mathbf{a}_t^{(i)}\}_{i=1}^K$ for time step t, different aggregation strategies can be defined as follows:

$$\text{Averaging:} \quad \mathbf{a}_t^{(\mu)} = \frac{1}{K}\sum_{i=1}^{K} \mathbf{a}_t^{(i)}, \tag{1.12}$$

$$\text{Majority Voting:} \quad \mathbf{a}_t^{(\text{MAX})} = \arg\max_{c\in\mathcal{C}} \sum_{i=1}^{K} \mathbf{1}\big(\mathbf{a}_t^{(i)} = c\big), \tag{1.13}$$

$$\text{EWE:} \quad \mathbf{a}_t^{(\text{EWE})} = \frac{1}{\sum_{i=1}^{K} r^{(i)}} \sum_{i=1}^{K} r^{(i)}\, \mathbf{a}_t^{(i)}, \tag{1.14}$$

where $\mathcal{C}$ denotes the set of possible affect labels, $\mathbf{1}(\cdot)$ is the class indicator function, and $r^{(i)}$ represents the reliability weight assigned to annotator i. In the case of continuous annotations, $\mathbf{a}_t^{(i)} \in \mathbb{R}$, averaging and EWE reduce to weighted linear combinations, while for categorical labels, majority voting is typically employed.

Nevertheless, all such quantification of the ground-truth — $\mathbf{a}_t^{(\mu)}$, $\mathbf{a}_t^{(\text{MAX})}$, and $\mathbf{a}_t^{(\text{EWE})}$ — ultimately collapses a diverse set of subjective annotations into a single estimate of central tendency. In doing so, these approaches neglect the inherent subjectivity and ambiguity of individual annotations *and impose hard, point-valued targets during training, encouraging neural networks to learn overconfident representations that fail to reflect underlying label uncertainty*, [186], [189], [190]. For example, when the annotations $\{\mathbf{a}_t^{(i)}\}_{i=1}^K$ follow a normal distribution, the average $\mathbf{a}_t^{(\mu)}$ provides a reasonable approximation of the consensus, since it reflects the symmetry of the distribution. However, when the distribution is skewed—often the case due to high subjectivity and ambiguity in the perception of affective expressions—these aggregation-based ground-truths may fail to capture the majority view. Although $\mathbf{a}_t^{(\text{EWE})}$ gives more influence to consistent annotators and resembles a weighted form of majority voting, it still disregards divergent perspectives and less consistent, yet potentially valid, interpretations of affect. In doing so, such methods tend to under-represent less frequent or extreme affective classes [191]. These outlier annotations, though rare, may encode subtle but important signals that are critical for a richer understanding of complex emotional dynamics, particularly in group interactions.

In light of these challenges, we strongly believe that reliable AER systems - particularly in real-world applications - must move beyond consensus-based modeling of ground-truth

labels to explicitly accounting for the inherent subjectivity in those labels [34], [192]. Doing so not only improves model robustness but also enables the integration of AER systems in human-in-the-loop solutions and supports the development of algorithms for active learning, co-training, and curriculum learning [191]. Motivated by this, prior research has attempted to address these challenges by introducing an additional ground-truth that explicitly captures the uncertainty in affect annotations. This ground-truth is typically represented as soft labels, in the form of either the standard deviation [189], [193], entropy-based measures [194], or by treating the annotations $\{\mathbf{a}_t^{(i)}\}_{i=1}^K$ as a distribution [195], [196]. The resulting representation can then be modeled within a multi-task learning (MTL) framework using auxiliary loss functions, such as a Kullback–Leibler divergence (KL) divergence loss term [191], [195] or a CCC based loss [189], [193] on these uncertainty representations and it's estimates.

In contrast to these works, in this thesis, to jointly model the uncertainty and the consensus in affect labels, we consider the ground-truth of affect to be a *distribution* of annotations. That is, we model the set of affective annotations as samples from a probability distribution:

$$\{\mathbf{a}_t^{(i)}\}_{i=1}^K \sim P, \qquad P \in \mathcal{P} \text{ arbitrary}, \tag{1.15}$$

where $\mathcal{P}$ denotes a generic family of distributions. This probabilistic approach provides a principled way to capture this uncertainty by representing annotations as probability distributions rather than single values, thereby preserving both central tendencies (e.g., the most likely affect) and dispersion (e.g., subjectivity and ambiguity in annotators' annotations).

Several approaches to distribution-based annotation modeling have been proposed in the AER literature. For instance, [195] trained a MTL network with a KL divergence loss by modeling affect annotations as a *univariate Gaussian*. In contrast, [197] represented label distributions using *histograms* to train their model. Similarly, [198] formulated emotional perception as a *multidimensional Gaussian* distribution, where the covariance structure captures dependencies between emotional categories, and used its mean as a soft label to train a deep neural network. In the broader field of ML and DNN research, uncertaintuy in data (e.g, noisy put data, ambiguous data labels) are often tackled under the appraoch of *Uncertainty Modeling* [199]–[201].

Uncertainty Modeling

Theoretical frameworks and modeling strategies for characterizing uncertainty in data are diverse and often tailored to the specific task and the nature of the data. Nevertheless, most studies distinguish between two principal types of uncertainty: *model uncertainty* (epistemic) and *data uncertainty* (aleatoric) [199], [202].

Model uncertainty (epistemic uncertainty) arises from limited knowledge or an insufficiently representative training corpus, and it manifests as uncertainty in the model parameters [202]. A model trained on a narrow or biased dataset is more prone to uncertainty when encountering inputs that deviate from the training distribution; in this sense, a larger and more diverse training corpus makes such out-of-distribution (OOD) samples less likely to occur during inference. This type of uncertainty is *reducible*—that is, it can be *explicitly modeled and quantified* and subsequently diminished—because it originates from incomplete knowledge about the task. In practice, epistemic uncertainty can be mitigated by expanding the diversity and coverage of the training data, thereby reducing the likelihood that the model encounters unfamiliar patterns. From a modeling perspective, epistemic uncertainty can be captured by treating model parameters probabilistically, representing them as distributions rather than

fixed point estimates. Such probabilistic formulations enable the quantification of uncertainty arising from limited knowledge and, as more data refines these parameter distributions, allow this uncertainty to be systematically reduced [199], [200].

In contrast, **data uncertainty** (aleatoric uncertainty) originates from inherent noise in the data itself, such as sensor measurement errors, modality-specific recording noise, or ambiguities in labels [202]. In image and video processing, this type of uncertainty is often linked to noisy labels or object occlusions [201], [203], while in speech processing, it may result from background noise or interfering speech [204], [205]. Unlike model uncertainty, data uncertainty is typically considered *irreducible*, since it stems from intrinsic randomness or noise in the data that cannot be eliminated even with additional training. Therefore, rather than reducing it, an effective strategy is to explicitly model and incorporate this uncertainty into the learning process. The task of affect recognition provides a representative example of *data uncertainty*, due to the inherent subjectivity and ambiguity in affective labels and annotations. Consequently, this thesis specifically focuses on *data uncertainty*, which we refer to as **label uncertainty**, as in our setting the uncertainty in the data primarily arises from the labels.

1.5.2 Probabilistic Modeling

As discussed earlier, *model uncertainty* can often be *modeled and quantified* through stochastic and probabilistic approaches. In particular, probabilistic modeling offers a principled framework for quantifying and capturing uncertainties during the learning process [199], [201], [202], [205]. Probabilistic modeling with DNNs refer to the paradigm where a neural network is used not only to predict a single deterministic output, but to parameterize a probability distribution over possible outcomes. Formally, given input features $\mathbf{X}$, a neural network with parameters $\mathbf{w}$ outputs the parameters $\phi(\mathbf{X};\mathbf{w})$ of a distribution $P_{\phi(\mathbf{X};\mathbf{w})} \in \mathcal{P}$. The predictive distribution is then written as

$$\mathcal{Y} \sim P_{\phi(\mathbf{X};\mathbf{w})}. \tag{1.16}$$

For example, in regression tasks, the network may output a mean $\mu(\mathbf{X};\mathbf{w})$ and variance $\sigma^2(\mathbf{X};\mathbf{w})$, defining a Gaussian likelihood $\mathcal{Y} \sim \mathcal{N}(\mu(\mathbf{X};\mathbf{w}), \sigma^2(\mathbf{X};\mathbf{w}))$. Under this formulation, the traditional, deterministic data-driven mapping of the affect recognition task (1.8), i.e., $\hat{\mathbf{a}}_t = f_{\mathbf{w}}(\mathbf{X}_{t_1:t_2})$, can be *reformulated* by treating the network output as the parameters of a predictive distribution rather than a single point. That is,

$$\mathbf{a}_t \sim P_\phi(\,\cdot \mid \mathbf{X}_{t_1:t_2};\, \mathbf{w})\,. \tag{1.17}$$

Note that, here, $\mathbf{w}$ denotes the parameters of the DNN model (e.g., network weights and biases), while ϕ represents the parameters of the output distribution (e.g., the mean $\mu_{\mathbf{w}}(\mathbf{X})$ and standard deviation $\sigma_{\mathbf{w}}(\mathbf{X})$, in case of a Gaussian assumption on the output distribution) predicted by the network. In this frequentist view, the parameters $\mathbf{w}$ are considered fixed after training, and the predictive distribution captures the label uncertainty through ϕ. In contrast, Bayesian inference provides a more general framework by treating the model parameters themselves as random variables. A prior distribution is placed over w,

$$\mathbf{w} \sim P(\mathbf{w}), \tag{1.18}$$

and learning corresponds to computing the posterior distribution given training data $\mathcal{D}$,

$$P(\mathbf{w} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathbf{w})\, P(\mathbf{w})}{P(\mathcal{D})}. \tag{1.19}$$

A BNN implements this idea by maintaining a distribution over the network parameters instead of a single point estimate. Predictions are obtained by marginalizing over the posterior:

$$P(\hat{\mathbf{a}}_t \mid \mathbf{X}_{t_1:t_2}, \mathcal{D}) = \int P_{\phi(\mathbf{X}_{t_1:t_2};\mathbf{w})}\; P(\mathbf{w} \mid \mathcal{D})\, dw. \tag{1.20}$$

This formulation unifies both sources of uncertainty: model uncertainty, captured by the posterior distribution over $\mathbf{w}$, and label uncertainty, captured by the predictive distribution P_ϕ. In affect recognition, this allows models not only to predict the most likely affective state but also to quantify the confidence of these predictions, accounting for both uncertainty in annotations $\{\mathbf{a}_t^{(i)}\}_{i=1}^K$ and limited data. In neural networks with millions of parameters, deriving the exact posterior distribution $P(\mathbf{w} \mid \mathcal{D})$ is not feasible, since computing the evidence $P(D)$ involves solving high-dimensional integrals that cannot be evaluated analytically. To address this challenge, the literature has introduced a variety of approximation methods.

Estimation of Model Uncertainty via stochastic model parameters $P(\mathrm{w} \mid \mathcal{D})$

A variety of approaches have been proposed to implement BNNs, differing mainly in how they *approximate the posterior distribution* over network parameters. Two such prominent approaches are: (1) Monte Carlo (MC) Dropout [202] and (2) Bayes by backpropagation (BBB) [200]. The MC Dropout approach [202] leverages the application of dropout at inference time as an approximate variational inference. By performing multiple stochastic forward passes through the network, MC Dropout effectively samples from an approximate posterior over the model's weights. This provides a simple yet powerful way to quantify epistemic uncertainty while requiring minimal modification to standard neural network architectures. MC Dropout and BBB share a common foundation in *variational inference*. Both approaches approximate the intractable true posterior $P(\mathbf{w} \mid \mathcal{D})$ with a tractable variational distribution $q(\mathbf{w} \mid \theta)$, and optimize this approximation by minimizing the KL divergence. The key difference is that MC Dropout imposes a Bernoulli variational distribution (via randomly dropping units), while BBB explicitly learns a parametric distribution (typically Gaussian) over each weight.

In BBB [200], each network parameter w is treated as a random variable, typically modeled with a Gaussian variational posterior $q(\mathbf{w} \mid \theta)$, where θ represents the variational parameters (e.g., mean $\mu_\mathbf{w}$ and variance $\sigma_\mathbf{w}^2$). To ensure non-negativity of $\sigma_\mathbf{w}$, the standard deviation is reparameterized as $\sigma_\mathbf{w} = \log(1 + \exp(\rho_\mathbf{w}))$. Hence, the variational parameters can be written as $\theta = (\mu_\mathbf{w}, \rho_\mathbf{w})$ and optimized through standard backpropagation, making training in BBB similar to that of a conventional neural network.

The variational posterior is optimized by minimizing the KL-divergence between $q(\mathbf{w} \mid \theta)$ and the true Bayesian posterior, yielding the negative evidence lower bound (ELBO):

$$f(\mathbf{w}, \theta)_{\mathrm{BBB}} = \mathrm{KL}[q(\mathbf{w}|\theta) \| P(\mathbf{w})] - \mathbb{E}_{q(\mathbf{w}|\theta)}\big[\log P(D \mid \mathbf{w})\big]. \tag{1.21}$$

Stochastic predictions in BBB are obtained via multiple forward passes (n), where weights w are sampled from $q(\mathbf{w}|\theta)$ at each pass, producing n stochastic estimates of $\widehat{\mathcal{Y}}_t$. Incorporating

this stochasticity, (1.21) is approximated as

$$\mathcal{L}_{\text{BBB}} \approx \sum_{i=1}^{n} \log q(\mathbf{w}^{(i)}|\theta) - \log P(\mathbf{w}^{(i)}) - \log P(D|w^{(i)}), \tag{1.22}$$

where $w^{(i)}$ denotes the i^{th} sample from $q(\mathbf{w}|\theta)$. The reparameterization trick [59] is used to enable efficient stochastic gradient descent by expressing random weight samples as deterministic functions of parameters and noise. For time-continuous AER, the BBB window size b controls how frequently new weights are sampled, with uncertainty assumed constant within each window. Unlike MC dropout, which samples implicit subnetworks, BBB explicitly learns distributions over weights, thereby providing a principled mechanism to capture both epistemic uncertainty (arising from the model) and aleatoric uncertainty (arising from the data). In this way, the uncertainty introduced during optimization—via the reparameterization of weight distributions—contributes to both sources of uncertainty.

Existing literature highlights that clearly distinguishing model (epistemic) and data (aleatoric) uncertainty is challenging, as the boundary between the two is often task- and context-dependent [205], [206]. In practice, data uncertainty *may* appear reducible through additional data or by reformulating the problem, which complicates its traditional interpretation as irreducible noise. For example, increasing the feature dimensionality or improving the quality of input features can make previously ambiguous samples more separable, thereby reducing the observed ambiguity in the data; however, this may simultaneously increase epistemic uncertainty because the model must learn a more complex mapping. These observations highlight that model and data uncertainty are not strictly separable, and changes in data representation or problem formulation can shift uncertainty from one category to the other.

However, given the difficulty of separating different uncertainty types, it is more pragmatic to concentrate on how uncertainty is modeled within the scope of a specific task. In the case of affect recognition—and in line with one of the central goals of this thesis, namely addressing the subjectivity and ambiguity of affect—this focus translates into modeling *label uncertainty*. A principled way to capture it is through the choice of the output distribution family P_ϕ, as discussed next. While prior work in AER has explored probabilistic modeling to jointly capture model and label uncertainty (e.g., [186], [196]), the explicit treatment of *label uncertainty* alone has received little attention. This lack of focused approaches represents a significant gap in the literature, despite the central role of subjectivity and annotator disagreement in affective perception.

Estimation of Label Uncertainty via choice of output distribution P_ϕ

The subjective and ambiguous nature of affective annotations naturally gives rise to a *ground-truth distribution* of labels, reflecting the inherent variability and uncertainty in how different annotators perceive affective expressions [191]. In line with the probabilistic modeling approach introduced earlier, *variational inference* methods such as BBB and MC dropout enable stochastic predictions through multiple forward passes of the network, effectively forming a *distribution over estimates* [200], [202]. In particular, the BBB approach achieves this by sampling different weight configurations from an optimized Gaussian posterior, resulting in variability across forward passes.

To explicitly estimate label uncertainty, our objective is therefore twofold: (i) optimize the parameters of the weight distribution, $P(\mathbf{w} \mid \mathcal{D})$, to produce stochastic outputs (capturing *model uncertainty*) as in (1.21), and (ii) simultaneously optimize the parameters of the output

distribution, P_ϕ, so that it directly models the annotation distribution $\{\mathbf{a}_t^{(i)}\}_{i=1}^K$. Unlike tasks such as uncertainty-based speech enhancement [205] or image classification [201], where label uncertainty is only implicitly captured through ELBO-based optimization of the weight parameters (1.21), in affect recognition we have access to explicit ground-truth distributions of annotations. This availability motivates us to go beyond implicit modeling and directly capture label uncertainty by leveraging the annotation distributions. Such treatment of the ground-truth as a distribution is typically studied in DNN research under the topic of *LDL*. LDL is a paradigm where each sample ground-truth is described by a *distribution* over multiple labels rather than a single label, capturing the degree to which each label applies [207]. This makes LDL well-suited for tasks with subjective or ambiguous annotations, such as affect recognition.

In the fields of AER and SER, BNN-based approaches for capturing model uncertainty [186], [191], [196] and LDL-based approaches for capturing label uncertainty [189], [190], [193], [195], [197] have largely been studied in isolation. However, their combination is arguably more intuitive for modeling subjective or ambiguous affect annotations, while also offering a more principled and statistically grounded framework for representing label uncertainty in affect recognition. In this thesis, we combine the advantages of BNN and LDL to address *label uncertainty* in affect recognition.

Choice of Distribution Family for the Output Distribution P_ϕ In the relatively limited prior work on this topic [190], [195], [197], the true distribution of affect annotations is generally unknown, leading most approaches to adopt simplifying assumptions—most often modeling annotations using either a Gaussian distribution or a histogram-based representation. The Gaussian distribution,

$$\mathcal{Y}_t \sim \mathcal{N}(\mu, \sigma), \tag{1.23}$$

$$p(y \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}\left(\tfrac{y-\mu}{\sigma}\right)^2\right], \tag{1.24}$$

is parameterized by mean μ and standard deviation σ, and provides a continuous probability model for real-valued random variables. Owing to its simplicity and tractability, the Gaussian has become the most common assumption for modeling affect annotations in label distribution learning (e.g., [195], [200], [208]).

A key limitation of the Gaussian distribution, however, is it's limited ability to model outliers—an issue that becomes particularly pronounced when data are *limited* or *sparse*, as is often the case in affective annotation [184], with typically ≤ 6 annotators. In such scenarios, the Student's t-distribution has been shown to provide a more robust and realistic alternative [184], [209]. Publicly available affect recognition datasets typically contain only a small number of annotations (e.g., 3–6 per sample [42], [115], [156], [210]), and collecting more annotations is both costly and resource-intensive [191], [211]. At the same time, the inherently subjective nature of affect perception further increases the likelihood of annotation outliers [35], [186].

The Student's t-distribution directly addresses these issues. It accounts for uncertainty in variance estimation and is particularly well-suited for small sample sizes or datasets with

outliers [209], [212]. Its density is defined as

$$\mathcal{Y}_t \sim \mathcal{N}(\nu, \mu, \sigma), \tag{1.25}$$

$$p(y \mid \nu, \mu, \sigma) = \frac{1}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\sqrt{\nu\sigma^2}} \left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}, \tag{1.26}$$

$$\text{where} \quad B(i,j) = \frac{\Gamma(i)\,\Gamma(j)}{\Gamma(i+j)}, \tag{1.27}$$

with location parameter μ, scale parameter σ, and degrees of freedom ν. The shape of the t-distribution resembles that of the Gaussian but features heavier tails, allowing it to better accommodate outliers. The degrees of freedom ν govern how close the distribution is to Gaussian—larger values of ν yield distributions increasingly similar to $\mathcal{N}(\mu, \sigma^2)$ [213]. Mathematically, as $\nu \to \infty$,

$$\left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \longrightarrow \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \tag{1.28}$$

and using the asymptotic property of the Beta function,

$$B\left(\tfrac{1}{2}, \tfrac{\nu}{2}\right) = \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)} \xrightarrow[\nu\to\infty]{} \sqrt{2\pi}\,\nu^{-1/2}. \tag{1.29}$$

Substituting these limits, the t-density simplifies to

$$p(y \mid \nu, \mu, \sigma) \xrightarrow[\nu\to\infty]{} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) = \mathcal{N}(y \mid \mu, \sigma^2). \tag{1.30}$$

Hence, the Student's t-distribution converges to the Gaussian distribution as $\nu \to \infty$, with the tail heaviness diminishing accordingly.

By explicitly incorporating the number of observations through the degrees of freedom, the t-distribution adapts to the sparsity of annotations and provides more robust estimates of central tendency and variability in the presence of outliers. This robustness, particularly under conditions of *limited* and *sparse* annotations typical of affect recognition datasets, constitutes the key motivation for adopting the t-distribution to model affect annotations in this work.

Label Distribution Learning Loss Function A widely used loss function in LDL is the KL-divergence [214]. This loss, denoted as $\mathcal{L}_{\text{KL}}$, measures the divergence between the ground-truth affect distribution $\mathcal{Y}$ and the estimated annotation distribution $\widehat{\mathcal{Y}}_t$ obtained from the model's stochastic outputs $\widehat{y}_t$. A smaller KL divergence (closer to 0) indicates higher similarity between the distributions, while a larger value implies greater dissimilarity. Minimizing this loss encourages the model to produce output distributions that align closely with subjectively annotated emotion distributions, thereby effectively capturing label uncertainty.

For a Gaussian assumption on $\mathcal{Y}_t$ (1.23), the KL divergence between two Gaussians, $\mathcal{Y}_t \sim$

$\mathcal{N}(m_t, s_t^2)$ and $\widehat{\mathcal{Y}}_t \sim \mathcal{N}(\widehat{m}_t, \widehat{s}_t^2)$, is given by[7]:

$$\mathcal{L}_{KL}(\mathcal{Y}_t \| \widehat{\mathcal{Y}}_t) = \log\left(\frac{\widehat{s}_t}{s_t}\right) + \frac{s_t^2 + (m_t - \widehat{m}_t)^2}{2\widehat{s}_t^2} - \frac{1}{2}. \tag{1.31}$$

Since KL divergence is asymmetric, the order of distributions matters. In Eq. (1.31), we specifically use $\mathcal{L}_{KL}(\mathcal{Y}_t \| \widehat{\mathcal{Y}}_t)$, i.e., the true distribution $\mathcal{Y}_t$ as the reference and $\widehat{\mathcal{Y}}_t$ as the approximating distribution. Minimizing this KL divergence mathematically enforces the approximating Gaussian to match the true mean and variance:

$$\frac{\partial \mathcal{L}_{KL}}{\partial \widehat{m}_t} = -\frac{m_t - \widehat{m}_t}{\widehat{s}_t^2} = 0 \quad \Rightarrow \quad \widehat{m}_t = m_t,$$
$$\frac{\partial \mathcal{L}_{KL}}{\partial \widehat{s}_t} = \frac{1}{\widehat{s}_t} - \frac{s_t^2}{\widehat{s}_t^3} = 0 \quad \Rightarrow \quad \widehat{s}_t^2 = s_t^2.$$

Thus, minimizing $\mathcal{L}_{KL}(\mathcal{Y}_t \| \widehat{\mathcal{Y}}_t)$ corresponds to a *mean-seeking* approximation rather than a mode-seeking one, capturing the full distribution [215, p. 76].

For a t-distribution assumption on $\mathcal{Y}_t$ (1.25), we derive the KL divergence between $\mathcal{Y}_t \sim t_\nu(m_t, s_t^2)$ and Gaussian outputs $\widehat{\mathcal{Y}}_t \sim \mathcal{N}(\widehat{m}_t, \widehat{s}_t^2)$. The loss is formulated as (see [P1] for the complete derivation):

$$\mathcal{L}_{KL} = \tfrac{1}{2}\log(2\pi\widehat{s}_t^2) + \frac{s_t^2 + (m_t - \widehat{m}_t)^2}{2\widehat{s}_t^2} - H(\mathcal{Y}_t). \tag{1.32}$$

Assuming a Gaussian for $\widehat{\mathcal{Y}}$ is justified, since the number of stochastic samples n used to approximate $\widehat{\mathcal{Y}}$ can be controlled. In this work, we set $n \geq 30$, ensuring that the t-distribution converges toward a Gaussian [213], [216]. A further benefit of this assumption is that the KL divergence simplifies to a tractable form between a Gaussian and a t-distribution, whereas computing the KL divergence between two t-distributions would require intractable expectations.

Thesis Contributions 1

In this thesis, we address label uncertainty—arising from the subjectivity and ambiguity of affective annotations—by integrating BBB-based probabilistic modeling with LDL-based estimation of label uncertainty. As shown in Figure 1.1, this constitutes the *first component* of the **perception and analysis aspect of social intelligence** in group interactions and is formulated as research questions **RQ1** and **RQ2** in Section 1.7. The corresponding publications are [P1]–[P3].

1.5.3 Affect at the Group-level [P4]

Another challenge in researching AER concerns a frequently overlooked characteristic of affect—its multilevel nature. As illustrated in Figure 1.1, the *recognition of the group-level affect* contributes towards the *perception* aspect of social intelligence in group interactions.

[7] The derivation and proof of this formula can be seen in Soch, Joram, et al. (2024). *StatProofBook/StatProofBook.github.io: The Book of Statistical Proofs (Version 2023)*. Zenodo. https://doi.org/10.5281/ZENODO.4305949, see also https://statproofbook.github.io/P/norm-kl.html

The relevance of collective group affect towards the functioning and outcomes of purposive groups is documented well in the literature [43], [49], [91], [217]–[220]. For example, a widely cited review of the literature discusses the importance of group affect for shaping (i) group member attitudes, (ii) cooperation and conflict resolution, (iii) creativity and decision making, and (iv) group effectiveness and performance [17].

Affect does not exist solely at the individual level; it also manifests as a shared construct within group interactions. Importantly, there are both conceptual and empirical distinctions between the affective experiences of individual group members and the collective mood of the group as a whole. These differences extend not only to how they are measured but also to how they influence group outcomes [221]. Group affect can therefore be understood as a collective social construct, reflecting the shared mood or emotions experienced jointly by group members at a given point in time [17], [89]. Barsade and Knight [17] formally define group affect as:

> ***collective affective state of the group, which is the amalgamation of group members' affective states expressed during group interactions (i.e., a bottom-up process) and which in turn affects future affective experiences of the group (i.e., top-down influence).***

Furthermore, scholars have emphasized the temporal dynamics of group affect, illustrating how momentary affective experiences—both individual and collective—serve as inputs that shape subsequent group-level affective states [222], [223].

Datasets for studying Group Affect

Group affect in this thesis is formulated as a *dynamic* and *multilevel* construct. Within this formulation of group affect, that is consistent with decades of affect research in organizational psychology literature [17], [43], [89], [223]–[225], a review on group affect, spanning the literature on organizational psychology, management, and computer science, and identified *two* key shortcomings and associated challenges.

First, the limited literature on group affect to date has predominantly relied on annotations of static images [49], [226]–[228] or temporally independent video segments [220] *without accounting for any temporal context*—temporal dynamics and change [43]. There is a feasibility argument to be made for both of these cases. Static images are much easier to obtain compared to observational group interaction data. Similarly, annotating video segments in an temporally independent manner without the temporal context is resource efficient, where annotators need not continuously track the group's affective expressions and its associated changes between consecutive segments. However, such an annotation approach is agnostic to the inherently dynamic nature of group affect [17], [43], [91], which limits the resulting empirical contributions and represents a misalignment between the theoretical construct of collective group affect as a dynamic process and its measurement [73], [219].

Second, from a GER standpoint, there is *theory-method misalignment* between conceptual approaches and theoretical models of group affect as presented in the organizational behavior literature (e.g., [17], [89]) and the group affect recognition methodologies developed in computer science studies such as [49], [227], [228]. While the theory conceptualizes group affect in complex *purposive group settings* involving interpersonal relationships and dynamics [17], group affect recognition research has focused on simpler groups that lack social intactness and interdependence towards a shared goal. As one example, consider the groups captured in the

Video Group Affect dataset, which includes concerts, silent protests, and fights [220]. Notably, annotating and modeling collective group affect in purposive groups is significantly more complex, as it requires capturing the dynamics of interpersonal relationships, unlike simpler non-purposive groups, where such relationships are largely absent and do not necessitate its modeling.

The above two challenges have contributed to the lack of publicly available GER datasets that capture group affect as a dynamic and multidimensional construct emerging in purposive group interactions. See Table 1.3 and Section 1.4.3 for further discussion. In this thesis, we address this limitation by leveraging the Memory Aware Conversational AI (MEMO) corpus [157], [229] of group interactions. To bridge the gap in datasets suitable for studying group affect in line with the properties discussed above, we developed an annotation strategy that not only tackles the key challenges of group affect annotation but also ensures methodological consistency with theoretical frameworks from organizational psychology [17], [225], [230]. A central focus of our approach is the challenge of capturing the temporal dynamics inherent in group affect. To this end, we employed an iteratively tuned 15-second window that enables a more precise representation of how group affect evolves and fluctuates over time. Moreover, unlike prior studies [49], [167], [228], [231] that have primarily examined non-purposive groups lacking social intactness and goal-oriented interdependence, our work emphasizes complex purposive groups characterized by dynamic interpersonal interactions. To address the complexity of annotating affect in such groups, we trained organizational psychology students extensively, ensuring that the resulting labels were context-aware and of high quality.

Group-level modeling

As introduced earlier in (1.4), the modeling of collective group affect for the task of GER in this thesis is formulated as follows, restated here for clarity:

$$P_{\mathbf{w}}\Big(h_{\psi}\Big(\{\mathbf{X}_{t_1:t_2}^{(i)}\}_{i=1}^{M} \Big) , \mathbf{a}_t^{(g)} \Big),$$

where w denotes the learnable weight parameters, analogous to the individual-level affect recognition case in (1.8), and h_ψ represents the additional parameters used to model the relationships between lower-level individual behavioral cues, interpersonal dynamics, and the corresponding emergent group-level affect. The choice of $h_\psi(\cdot)$ can vary in complexity and sophistication. It may range from simple pooling and aggregation operations to more advanced mechanisms such as attention-based models [232] or graph-based approaches [233]. While simple operations may not require training learnable parameters ψ, more sophisticated approaches typically do. Between these two extremes lie synchrony- and mimicry-based interpersonal features [234], which provide an intermediate level of modeling group affect. A detailed discussion of these three levels of modeling is presented in the following paragraphs.

Aggregation measures The most basic form of group-level modeling is the aggregation-based approach. Here, individual-level behavioral cue features or representations extracted from AER models are combined into group-level features and representations using statistical heuristics such as `mean`, `mode`, `median`, `standard-deviation`, `minimum`, and `maximum`. This process can be formalized as:

$$h_{\psi}\Big(\{\mathbf{X}_{t_1:t_2}^{(i)}\}_{i=1}^{M} \Big) , \tag{1.33}$$

where ψ denotes the non–data-driven statistical heuristics. The resulting aggregated features can then be combined with learnable parameters w to predict the ground-truth of group-level affect $\mathbf{a}_t^{(i)}$, as in (1.4). This approach views group-level affect as a straightforward aggregation

of individual-level affect, thereby capturing only the *bottom-up* processes. However, the formal definition of group affect, presented in 1.5.3, also highlights the role of *top-down* dynamic processes, which require modeling interpersonal relationships. A key limitation of the aggregation approach is that it neglects such interactions—for instance, statistical dependencies between the cues and representations of different group members, such as between interactant 1: $\mathbf{X}^{(1)}_{t_1:t_2}$ and interactant 2: $\mathbf{X}^{(2)}_{t_1:t_2}$.

Interpersonal measures One way to capture interpersonal relationships for modeling group-level constructs such as group affect is through handcrafted interpersonal features like *synchrony* and *mimicry* [40], [234], [235]. Mimicry refers to the often unconscious imitation of another person's verbal and nonverbal behaviors during social interactions [236]. This includes the replication of postures, gestures, facial expressions, tone of voice, speech rate, or even verbal phrasing. The process is typically automatic and serves important social functions such as empathy, rapport-building, and emotional bonding [236], [237]. Synchrony, in contrast, emphasizes timing rather than the specific nature of behaviors. Bernieri et al. [238] define synchrony as "the degree to which the behaviors in an interaction are non-random, patterned, or synchronized in both form and timing." In simpler terms, synchrony refers to the adaptation of one individual to the rhythms and movements of their partner [239], as well as the congruence between cycles of engagement and disengagement in interaction.

Both *synchrony* and *mimicry* have been empirically linked to higher-level constructs such as conversation quality [40], rapport [237], cohesion [235], and liking [240]. While multiple computational models exist to quantify them in group interactions [241], [242], a widely adopted formulation for extracting synchrony and mimicry features between two interactants—say interactant 1: $\mathbf{X}^{(1)}_{t_1:t_2}$ (or simply $\mathbf{X}^{(1)}$) and interactant 2: $\mathbf{X}^{(2)}_{t_1:t_2}$ (or simply $\mathbf{X}^{(2)}$)—is as follows:

Synchrony Features

$$\text{Correlation coefficient } \rho\text{: } \mathbf{X}^{(1)} \circledast \mathbf{X}^{(2)}$$

$$\text{Lagged correlation } \rho_\delta\text{: } \max z(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$$

$$\text{Best lag } \delta\text{: } \arg\max_l z(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, l) - ||\mathbf{X}^{(1)}|| + 1$$

$$z(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, l) = \sum_{k=0}^{||\mathbf{X}^{(1)}||-1} \mathbf{X}^{(1)}_l \circledast \mathbf{X}^{(2)}_{k-l+N-1}$$

Mimicry Features

$$\text{Global Mimicry: } \sum_{i=0}^{||\mathbf{X}^{(1)}||/2} (\mathbf{X}^{(1)}_i - \mathbf{X}^{(2)}_i)^2 - \sum_{j=||\mathbf{X}^{(1)}||/2}^{||\mathbf{X}^{(1)}||} (\mathbf{X}^{(1)}_j - \mathbf{X}^{(2)}_j)^2$$

$$\text{Symmetric Mimicry: } (\mathbf{X}^{(1)}_l - \mathbf{X}^{(2)}_l)^2 \circledast L$$

$$\text{Asymmetric Mimicry: } P(\mathbf{X}^{(2)}_b \mid \Theta_{\mathbf{X}^{(1)}_a}) \circledast L$$

Each of the above feature can represent the h_ψ in (1.4). Here, $z(\cdot, \cdot)$ denotes the cross-correlation function, $\circledast$ represents linear correlation between two signals, and $||\mathbf{X}^{(1)}||$ is the length of signal $\mathbf{X}^{(1)}$. The possible time lags are given by $l = 0, 1, \ldots, ||\mathbf{X}^{(1)}|| + ||\mathbf{X}^{(2)}|| - 2$,

with $N = \max(||\mathbf{X}^{(1)}||, ||\mathbf{X}^{(2)}||)$. Further, $L = [0, 1, \ldots, ||\mathbf{X}^{(1)}||]$ with $l \in L$, and Θ is the parameter of a Gaussian mixture model (GMM) trained via expectation–maximization on data points from $\mathbf{X}^{(1)}$ during the initial segment of the interaction ($a \in [0, m], m = 2 \cdot ||\mathbf{X}^{(1)}||/3$). Correspondingly, $\mathbf{X}_b^{(2)}$ represents the data points from $\mathbf{X}^{(2)}$ in the later segment of the interaction ($a \in [m, ||\mathbf{X}^{(2)}||]$).

While these interpersonal features represent an important step toward capturing group-level social constructs, *three* key limitations can be noted: (1) they are constrained to dyads within a group, meaning they exclusively capture synchrony and mimicry between pairs of interactants rather than accommodating an arbitrary number of participants. As a result, an additional aggregation strategy is required to combine dyad-level features into group-level representations. Here, the aggregations describe the distribution of dyadic-level features within a group thereby capturing the interpersonal nuances within all possible dyads in the group. The `average` and `standard-deviation` explains the average and the deviation from the average of synchrony measures across all possible dyads in a group. (2) they are not data-driven; in other words, h_ψ does not involve trainable parameters and thus cannot directly support data-driven modeling even when a dataset of group interactions is available. (3) this approach scales poorly with larger group sizes, since the number of dyads and their corresponding representations grows rapidly as the group size increases. As a result, data-driven modeling becomes challenging, as these representations are difficult to generalize across arbitrary group sizes and variable input lengths.

In this thesis, we move beyond handcrafted features to model interpersonal relationships with the aim of achieving a more data-driven representation of both interpersonal dynamics and group-level aggregations. To this end, our modeling approach draws on social network theory [243], where individuals are represented as nodes and their relationships as edges, enabling the model to learn both the structure and temporal evolution of social interactions. This approach is described in detail in the following section.

1.5.4 Graph Modeling

Grounded in social network theory [243], we frame the task of GER as a graph-based modeling problem using GNNs. In this formulation, a group interaction is represented as an undirected graph $\mathcal{G} = (V, E)$, where V denotes the set of M nodes (interactants) and E represents the set of edges (relationships between them). Each node $V_i \in V$ represents an interlocutor with individual-level features $X^{(i)} \in \mathbb{R}^F$, and each edge $E_{i,j} \in E$ denotes a connection between two nodes. The adjacency matrix A captures the edge structure, where $A_{i,j} = 1$ if nodes i and j are connected and $A_{i,j} = 0$ otherwise.

The overall task of GER can be formulated as a two-stage sequential process: first, group-level representations are constructed through a graph encoder parameterized by ψ, and second, these representations are mapped to the group-level ground-truth $\mathbf{a}_t^{(g)}$ using parameters $\mathbf{w}$. In this formulation, the aggregation defined in (1.4) and (1.33) becomes fully data-driven via the graph representation $\mathcal{G}_\psi$:

$$\mathcal{G}_\psi = h_\psi\Big(\{\mathbf{X}_{t_1:t_2}^{(i)}\}_{i=1}^{M} \Big), \tag{1.34}$$

which is then used by a parameterized regressor $f_{\mathbf{w}}$ (in our case, a MLP) to map the learned

graph features to group-level affect annotations:

$$\mathbf{a}_t^{(g)} = f_{\mathbf{w}}\Big(\mathcal{G}_\psi\Big). \tag{1.35}$$

A standard framework to train a graph neural networks (GNN), i.e., $\mathcal{G}_\psi$ is introduced by Kipf et al. [233], as graph convolution networks (GCN). This setup consists of two main steps: (1) *convolution*, where each node transforms its feature representation ($\mathbf{X}^{(i)}$) to share with adjacent nodes, and (2) *message passing*, wherein these features are propagated to the adjacent nodes. Nodes subsequently update their representations by aggregating information from their neighbors, typically using simple operations such as summation or averaging. However, this approach can produce identical output features for nodes with identical neighborhoods, thereby limiting the model's expressiveness for certain graph structures. For example, this poses a challenge in group affect modeling, where interlocutors are connected through a fully connected graph that captures overall group membership but overlooks the distinct relationships between individuals. To address this limitation, graph attention networks (GAT) offer a more data-driven, attention-based message passing mechanism [244].

Unlike basic sum or average aggregations in GCN, the attention-based GAT compute a weighted average of multiple node features. These weights are dynamically determined based on a combination of the node's own features, the interlocutor's individual-level features ($\mathbf{X}^{(i)}$), and the features of adjacent nodes ($\mathbf{X}^{(j)}$), which represent the interacting counterparts of the interlocutor. The GAT layer obtains an attention based aggregation formulated as:

$$\mathbf{X}_{(l+1)}^{(i)} = \Phi\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij}\mathbf{w}_{(l)}\mathbf{X}_{(l)}^{(j)}\right), \tag{1.36}$$

where $\mathbf{w}^{(l)}$ is the learnable weight parameters, at training iteration l, that transforms the node features, Φ represents an arbitrary activation function, and α_{ij}, the learnable attention weight between nodes i and j. The attention weight α_{ij} is formulated as:

$$\alpha_{ij} = \frac{\exp\Big(\text{LeakyReLU}\Big(\vec{\mathbf{a}}\Big[\mathbf{w}\mathbf{X}^{(l)}||\mathbf{w}\mathbf{X}^{(j)}\Big]\Big)\Big)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}\left[\mathbf{w}\mathbf{X}^{(i)}||\mathbf{w}h_k\right]\right)\right)}, \tag{1.37}$$

where $||$ denotes concatenation, $\vec{\mathbf{a}}$ represents the attention mechanism, implemented as a single-layer feedforward neural network parameterized by a weight vector $\vec{\mathbf{a}} \in \mathbb{R}^{2F}$, LeakyReLU is the chosen activation function, and $\mathcal{N}_i$ denotes the indices of the adjacent nodes of node i. To accommodate varying group sizes (g), we fix the number of nodes M to the maximum group size in the dataset. For groups with fewer than M members, dummy nodes without edges are added, ensuring that $A_{x,j} = 0$ for such nodes ($g < x \leq M$). To capture the temporal dynamics across consecutive time segments, we model the fluctuations of affect between timesteps $\mathbf{a}_{t-1}$, $\mathbf{a}_t$, and $\mathbf{a}_{t+1}$ (after the graph modeling stage) using an long short-term memory (LSTM) layer. To ensure that gradients can propagate across consecutive segments, we adopt a *mini-batching* strategy in which each batch is composed of temporally ordered segments from the same interaction, thereby preserving temporal continuity. This batching strategy is widely used in end-to-end continuous emotion recognition methods [174], [175].

Thesis Contributions 2

Within this research on group-level affect, our contributions are two-fold: (1) we provide, for the first time in the literature, fine-grained temporal annotations of group-level affective expressions in purposive group interactions using an annotation strategy grounded in organizational psychology; and (2) we introduce a novel, data-driven GNN-based method for modeling interpersonal dynamics in the task of GER. As illustrated in Figure 1.1, this work represents the *second component* of the **perception and analysis aspect of social intelligence** in group interactions and is framed in this thesis as **RQ2** and **RQ3** in Section 1.7. The corresponding publication is [P4].

1.6 Synthesis of Affect

As individuals' inner affective and emotional states manifest through outward expressions, both speech and visual cues have become central to research on *affect synthesis*, spanning multiple modalities and tasks. Within the visual modality, this has led to research on *gesture synthesis* [245], [246], *facial expression synthesis* [247]–[249], and *body-skeleton synthesis* [250], [251]. These efforts are predominantly framed as *co-speech* problems—i.e., generating gestures and expressions conditioned on corresponding speech. This trend highlights that, within the broader affect synthesis landscape, *speech* has consistently been recognized as the most fundamental modality in communication, particularly in dyadic and group interactions (as discussed in Section 1.4.1). While affective visual cue generation primarily aims to complement speech—making communication more natural, human-like, and socially engaging—speech remains the central channel through which both embodied and non-embodied agents interact [53], [55].

Acknowledging this, the present thesis focuses on *affect synthesis in the speech modality* as the first and most fundamental step toward enabling socially intelligent agent communication in group interactions. The generation of speech, commonly referred to as *speech synthesis* or text-to-speech synthesis (TTS), has been an active research field for decades, with early efforts exploring concatenative and statistical parametric approaches [252], [253]. Over the past decade, however, the field has undergone a paradigm shift: the landmark work of Oord et al. [254] revolutionized TTS, inaugurating the deep learning era and establishing a data-driven framework in which mappings from text to speech are learned directly from large datasets. Contemporary TTS models can now generate speech of remarkably high quality, often indistinguishable from human recordings. Yet, in the context of deploying TTS within socially intelligent agents for group interactions, high-fidelity synthesis alone is insufficient. Modern TTS systems are increasingly expected to produce *expressive speech* that conveys speaking styles and emotions, thereby approximating the natural variability and affective richness of human communication. To this end, applications in social intelligence demand the development of *emotion-controllable and emotion-conditioned speech synthesis*.

Within this scope, two complementary research directions have emerged: *emotion-conditioned speech synthesis* (ESS) [55], which seeks to directly generate speech from text input with specified emotional attributes, and *speech emotion conversion* (speech emotion conversion (SEC)) [58], which modifies existing speech to alter or impose emotional characteristics. Together, these paradigms form the foundation of affect synthesis in the speech modality and constitute the primary focus of this thesis. The distinction between these tasks lies largely in the type of conditional signal provided to the synthesis model. ESS typically takes

textual input together with an affective conditioning signal, whereas SEC instead operates on a reference speech signal—encoding linguistic content and speaker identity—along with an affective conditioning input to guide the transformation. These two research directions are often regarded as sister fields to TTS [255] and voice conversion (VC) [256], frequently borrowing and adapting methodologies and techniques across domains.

A key motivation for pursuing *speech emotion conversion* (SEC) rather than *emotional speech synthesis* (ESS) lies in its ability to preserve the naturalness of speech. Since SEC operates on an existing utterance, the linguistic content and speaker identity are inherently retained, while only the emotional characteristics are modified [58], [257]. This makes the converted speech perceptually more convincing than fully generated emotional speech, which often struggles to balance intelligibility, speaker identity, and emotional expressivity. Moreover, SEC is linguistically less complex than ESS, as it bypasses the need to generate speech directly from text and instead focuses solely on the affective transformation of an already intelligible signal [55]. This reduced complexity also makes SEC more data-efficient, since fewer parallel emotional corpora are required compared to training end-to-end expressive TTS systems. From an application standpoint, these advantages are particularly relevant to the deployment of embodied and non-embodied conversational agents in group interactions, where preserving naturalness and persona identity is crucial for maintaining social presence and engagement, and where emotional conversion of speech can enhance the perception of agents as socially intelligent communicators. To this, in this thesis, we specially focus on the task of SEC, which is formulated in the subsequent section.

Formulation: Affect synthesis in general can be defined as the counterpart to the unified formulation of affect modeling $P_{\mathbf{w}}(\mathbf{X}, \mathbf{a})$ (1.1), where the goal is to generate expressive behavioral signals that convey a desired target affect. From a probabilistic perspective, this corresponds to modeling the conditional distribution[8]

$$P_{\mathbf{w}}(\mathbf{X} \mid \mathbf{a}), \tag{1.38}$$

which describes the generation of behavioral cues $\mathbf{X}$ given a specified affect $\mathbf{a}$. In the case of SEC, this can be formulated as a deterministic mapping from a speech signal with a source affect $\mathbf{a}$ to a speech signal expressing a target affect $\mathbf{a}_{\text{trg}}$:

$$\hat{\mathbf{X}}_{\text{trg}} = f_\theta(\mathbf{X}, \mathbf{a}_{\text{trg}})\,, \tag{1.39}$$

where $\mathbf{X}$ denotes the source speech signal, $\mathbf{X}_{\text{trg}}$ the converted speech signal, and f_θ the parameterized DNN-based methodology. With this setup the task of SEC [55], [58], [258] can be defined as a sub-field of emotion-conditioned speech synthesis that:

aims to modify the emotion expressed in input speech while preserving lexical content and speaker identity.

This requires precise control over prosodic attributes that convey emotional content, such as intonation, stress, rhythm, and loudness, which are controlled by the acoustic features of speech sounds, such as fundamental frequency, duration, energy, and spectral envelope. The parameters $\mathbf{w}$ are then trained so that the generated output expresses the intended affect.

[8] In the synthesis of affect, we assume that the emotional state remains constant within a given speech segment. Accordingly, the overall affect level of the segment is denoted by a, without explicit reference to its temporal variation t. This simplified formulation is adopted throughout this section and in related discussions in the thesis. Importantly, this assumption does not substantially limit applicability, as temporal affect fluctuations can be addressed by synthesizing multiple shorter utterances conditioned on the target emotion.

Formally, this can be expressed as

$$\mathbf{w}^{\star} = \arg\min_{\mathbf{w}} \; \mathbb{E}\Big[\mathcal{L}\Big(\mathbf{X}_{\text{trg}}, \hat{\mathbf{X}}_{\text{trg}}\Big)\Big], \tag{1.40}$$

where $\mathcal{L}(\cdot, \cdot)$ measures the emotional discrepancy between the source and target speech. However, this setup and its associated loss function presuppose the availability of the ground-truth $\mathbf{X}_{\text{trg}}$, which the modeling function f_θ is trained to approximate.

1.6.1 In-the-wild and non-parallel speech data

The availability of ground-truth utterances $\mathbf{X}_{\text{trg}}$ is commonly referred to as parallel speech data. Such data consists of pairs where each source utterance $\mathbf{X}$ has a corresponding target utterance $\mathbf{X}_{\text{trg}}$ expressed in a different emotion. While widely used, parallel speech is highly resource-intensive to collect and scales poorly [55], [259], [260]. This limitation stems from the fact that parallel samples can only be obtained through scripted, acted-out sessions, where a speaker is instructed to deliver the same lexical content in multiple emotional styles. Beyond being impractical, this setup contradicts the view of affect as subjective, spontaneous, and context-dependent, as adopted in this thesis (see Section 3.1).

For these reasons, as the first in the literature of SEC and ESS, this thesis makes a deliberate shift away from traditional acted-out data towards in-the-wild speech. This direction not only aligns with psychological theories of affect and our formulation of affect but also offers clear methodological benefits. Unlike acted-out speech—which is essentially read aloud and often lacks natural variability—in-the-wild recordings capture spontaneous conversational dynamics, spontaneous emotional expressions, diverse speaking styles, nonverbal cues such as laughter or lip smacks, and disfluencies like hesitations or interruptions [42], [56], [57]. For instance, the in-the-wild MSP-Podcast dataset (v1.10) [42] provides substantially greater variability compared to predominant acted-out SEC datasets such as ESD [58] and IEMOCAP [187]. MSP-Podcast contains approximately 238 hours of audio, includes utterances of variable duration, features recordings from over 1,400 speakers, and captures naturalistic emotional expressions. In contrast, the ESD corpus offers only around 29 hours of acted-out speech from 10 English speakers, highlighting the limited scale and diversity of acted datasets relative to in-the-wild resources. Empirical analyses using the NaturalVoices dataset [57] confirm that models trained on such data generate speech that is both more natural and more intelligible. Crucially, this shift necessitates methods that can exploit in-the-wild data without relying on parallel speech samples. Despite the drawbacks of acted-out data, they remain widely used [258]–[264]—primarily because the availability of parallel samples enables simpler training strategies. Specifically, it allows for straightforward formulations such as the deterministic mapping in Eq.(1.39) and deterministic loss functions like Eq. (1.40). These simplifications make the task of SEC considerably easier to train and, importantly, facilitate evaluation by providing ground-truth target utterances $\mathbf{X}_{\text{trg}}$ against which system outputs can be directly compared.

Disentanglement of speech components In contrast to Eq. (1.39) and Eq. (1.40), techniques for SEC that do not rely on parallel speech samples must instead learn the following conditional probability:

$$P_{\mathbf{w}}(\mathbf{X} \mid \mathbf{a}), \tag{1.41}$$

$$\mathbf{X}_{\text{trg}} \sim P_{\mathbf{w}}(\mathbf{X} \mid \mathbf{a}_{\text{trg}}). \tag{1.42}$$

To model the conditional distribution in Eq. (1.41), a *disentanglement technique* is typically employed. The central idea is to decompose the input speech signal into distinct latent components, such as

$$\mathbf{z} = (\mathbf{z}_{\mathrm{l}},\ \mathbf{z}_{\mathrm{s}},\ \mathbf{z}_{\mathrm{a}}), \tag{1.43}$$

where $\mathbf{z}_{\mathrm{l}}$, $\mathbf{z}_{\mathrm{s}}$, and $\mathbf{z}_{\mathrm{a}}$ denote the latent representations of lexical content, speaker identity, and affect, respectively. Ideally, each latent representation exclusively capture the different speech attributes, allowing them to be manipulated independently. During inference, as formulated in Eq. (1.42), the affect embedding $\mathbf{z}_{\mathrm{a}}$ can be modified to reflect the target emotion while preserving $\mathbf{z}_{\mathrm{l}}$ and $\mathbf{z}_{\mathrm{s}}$. For example, the same speaker ($\mathbf{z}_{\mathrm{s}}$) uttering the same sentence ($\mathbf{z}_{\mathrm{l}}$) can be converted from a neutral style to an angry style simply by altering $\mathbf{z}_{\mathrm{a}}$. This enables emotion conversion without the need for parallel data.

Existing works have primarily employed variational auto-encoders (VAE) [265], [266] and sequence-to-sequence encoder-decoders [261], [264] to achieve disentanglement during training. However, more recently, SSL based representations have emerged, which have shown great success in several downstream tasks such as automatic speech recognition (ASR) [178], phoneme segmentation [267], speaker verification [180], and SER [37]. Existing literature well documents that representations obtained from SSL models finetuned on a specific task have exclusive information on that particular task. Analysis presented in [268] reveals that discrete representations obtained from HuBERT contain exclusively phoneme and lexical information, without any speaker or fundamental frequency information. In [180], the HuBERT framework [179] was used to build the WavLM which was then fine-tuned exclusively to preserve speaker identity [180]. Similarly, in [37], the `wav2vec` framework [178] is finetuned to capture the emotion information, thereby becoming the state-of-the-art in the task of SEC.

Disentangled representations (z): For the disentangled SSL-based representations, we use the following encoded features:

(i) *Lexical representation* ($\mathbf{z}_l \in \mathbb{N}^{1\times N}$): Following [258], [269], we adopt a HuBERT-based encoder E_l and use discrete HuBERT units obtained via k-means clustering on continuous HuBERT features. Formally, $\mathbf{z}_l = [z_1, \ldots, z_N]$, where each z_i is a positive integer and N is the length of the input discrete unit sequence, corresponding to the number of frames in HuBERT's representations. Prior studies [270], [271] have shown that these units strongly correlate with the phonemic content of the utterance. The feature rate of these speech units is 49 Hz.

(ii) *Speaker representation* ($\mathbf{z}_s \in \mathbb{R}^{512}$): Following [63], we extract a d-vector from a pre-trained WavLM-based speaker verification model E_s [180].

(iii) *Emotion representation* ($\mathbf{z}_a \in \mathbb{R}^{128}$): A continuous embedding is obtained by applying a trainable linear transformation to the emotion label $\mathbf{a}$ during training and to the target emotion label $\mathbf{a}_{\mathrm{trg}}$ during inference. Instead of a data-driven representation, we deliberately use this encoded label–based representation due to the lack of parallel samples at both training and inference. Although one could use a reference speech sample with the desired target affect, this approach risks leaking speaker information, since affect is often correlated with speaker identity. Such leakage could harm conditional generation and compromise speaker fidelity.

Notably, unlike $\mathbf{z}_l$, both $\mathbf{z}_s$ and $\mathbf{z}_e$ are global utterance-level representations. To align them with the frame-level $\mathbf{z}_l$ and to construct $\mathbf{z}$ (Eq. 1.43), we broadcast $\mathbf{z}_s$ and $\mathbf{z}_e$ across all frames/discrete units, yielding frame-level versions of these features.

Apart from disentangling the speech signal into its constituent latent representations, an additional component is required to perform conditional modeling and synthesis based on these representations. In this thesis, we focus on *generative modeling* approaches to realize this objective, as detailed in the following section.

1.6.2 Generative Modeling [P9]–[P11]

Generative models provide a principled framework for learning the conditional distributions (e.g., Eq. (1.41)) necessary for speech emotion conversion. Murphy [47], [48] define generative modeling as

the family of probabilistic methods that aim to model the joint distribution $P(\mathbf{x}, \mathbf{y})$ ***of data*** $\mathbf{x}$ ***and, optionally, associated labels*** $\mathbf{y}$***.***

By capturing this joint distribution, generative models allow us to perform a variety of tasks, such as sampling new data points $\mathbf{x} \sim P(\mathbf{x})$, imputing missing values, or evaluating likelihoods. This contrasts with discriminative models, which directly approximate the conditional distribution $P(\mathbf{y} \mid \mathbf{x})$ without modeling the data-generating process itself.

In the context of this thesis, generative models are employed to learn conditional distributions (i.e., Eq. (1.41)) that enable the synthesis of speech with controlled factors such as emotion, speaker identity, and lexical content (i.e., Eq. (1.42)). Within the generative modeling formulation, the synthesis based on the disentangled source input speech (Eq. (1.43)) can be reformulated as:

$$P_{\mathbf{w}}(\mathbf{X} \mid \mathbf{z}), \tag{1.44}$$

$$\mathbf{X}_{\text{trg}} \sim P_{\mathbf{w}}(\mathbf{X} \mid \mathbf{z}_{\text{trg}}, \mathbf{a}_{\text{trg}}). \tag{1.45}$$

A wide range of techniques have been developed for generative modeling, each with different assumptions and mechanisms for learning the underlying data distribution. Early approaches include *VAEs* [59], which introduce latent variables and optimize a variational bound on the data likelihood, and *GANs* [60], which frame generation as a minimax game between a generator and a discriminator. More recent advances include *normalizing flows* [272], which construct complex distributions by composing invertible transformations with tractable Jacobians[9], and *diffusion models* [61], which learn to reverse a noise corruption process through iterative denoising steps. These families of models have contributed to improvements across domains such as image [62], [185], text [65], and speech [64], [271], [273] generation.

In this thesis, GAN-based vocoders [270] and diffusion-based generative models are employed for the task of SEC, primarily due to its recent effectiveness in handling speech in tasks such as speech synthesis [64], [274], speech enhancement [275], [276], speech dereverberation [273], and speech editing [277], [278]. The following sections introduce GAN-based vocoder and diffusion models in greater detail and discuss their relevance for this thesis.

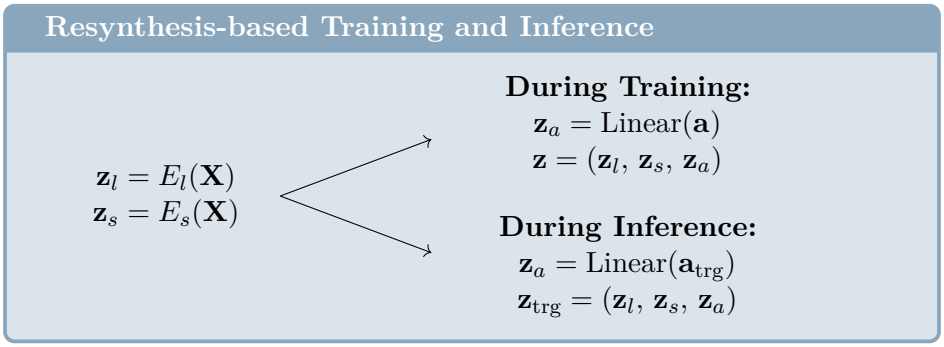
Neural Vocoder based generation [P9]

The *resynthesis*-based speech synthesis technique, Resynthesis with Visual Input for Speech REgeneration (ReVISE) [271], provides a direct and streamlined way to generate a speech waveform from speech tokens, which in our case correspond to disentangled representations.

[9]Here, the Jacobian refers to the matrix of partial derivatives of the transformation. In normalizing flows, the determinant of this Jacobian quantifies how probability mass is distorted under the transformation, and must be efficiently computable to enable exact likelihood evaluation.

For audio-visual speech regeneration, [271] formulate synthesis as a two-stage process: (i) predicting frame-level lexical tokens from a reference clean speech signal, and (ii) generating speech with a neural vocoder conditioned on these lexical tokens along with auxiliary cues such as lip movements and speaker identity. Although [271] do not explicitly frame their method in terms of representation disentanglement and emotion conversion, their approach nevertheless demonstrates how speech representations can be leveraged to perform controlled synthesis when conditioned on multiple modalities. Moreover, their approach, specifically in it's first stage, also uses a clean speech reference which in our case in analogous to the parallel emotional speech sample. In this thesis, we address these challenges to adapt the *resynthesis*-based neural vocoder for the task of SEC.

For the SEC, the ReVISE architecture is modified as follows: (i) in the first stage speech representations, including the frame-level lexical tokens $\mathbf{z}_l$, are extracted as disentangled representations as detailed in Section 1.6.1. Specifically, is formulated as:



Crucially, the key distinction between training and inference lies in how $\mathbf{z}_a$ is defined. Second (ii), a neural vocoder-based architecture is trained to reconstruct the source waveform $\mathbf{X}$ using the discrete representations $\mathbf{z}$. Here, during training, the need for parallel target samples $\mathbf{X}_{\text{trg}}$ is removed by reconstructing the input speech itself, i.e., *training to resynthesize the input speech* while conditioning on the disentangled speech units (Eq. 1.44). At inference time, only the emotion representation is replaced with that derived from the target affect label, $\text{Linear}(\mathbf{a}_{\text{trg}})$, thereby altering the expressed emotion in the generated speech while preserving the speaker identity and lexical content (Eq. 1.44). Throughout this thesis, we refer to this training–inference configuration as the *resynthesis* setup.

As the neural vocoder we use a modified version of the HiFiGAN [63], as introduced in [270]. Unlike the vanilla HiFiGAN which performs speech synthesis (i.e., speech waveform generation) based on an input Mel-spectrogram, the modified version [270] operates on the discrete speech representation such as the discrete HuBERT representation $\mathbf{z}_l$ (introduced in Section 1.6.1). Similar to the HiFiGAN, the modified architecture is also trained using an adversarial training setup. The architecture is trained using a generator G and two discriminator networks, the multi-*period* and the multi-*scale* discriminators, D_p and D_s, respectively [63], [270]. The generator G has a series of blocks composed of a transposed convolution and a dilated residual layer. The transposed convolutions upsample the representations to match the input number of samples T. The dilated residual layers increase the receptive field. The input to G is the concatenated disentangled representations $\mathbf{z}$ and the output is the resynthesised speech $\mathbf{X}$ conditioned on lexical, speaker and affect information. The multi-period discriminator D_p consists of six sub-discriminators each operating on different *period hops* of the input and generated speech: 2, 3, 4, 5, 7, and 11. Similarly, the multi-scale discriminator D_s uses three sub-discriminators operating at different *scales*: the original scale, 2x downsampled scale, and

4x downsampled scale.

For a resynthesised output speech signal $\hat{\mathbf{X}}$ each of the sub-discriminators D_j in D is trained to minimize the following adversarial losses ($\mathcal{L}_{adv}$ and $\mathcal{L}_D$):

$$\mathcal{L}_{adv}(D_j, G) = \sum_x ||1 - D_j(\hat{\mathbf{X}})||_2^2, \tag{1.46}$$

$$\mathcal{L}_D(D_j, G) = \sum_x [||1 - D_j(\mathbf{X})||_2^2 + ||D_j(\hat{\mathbf{X}})||_2^2], \tag{1.47}$$

where $\hat{\mathbf{X}} = G(\mathbf{z})$ is the resynthesised speech waveform based on the input speech representations $\mathbf{z}$. Along with $\mathcal{L}_{adv}$ and $\mathcal{L}_D$, a reconstruction loss term $\mathcal{L}_{recon}$ and a feature-matching loss $\mathcal{L}_{fm}$ [279] is also added to the loss function. $\mathcal{L}_{recon}$ measures the Mel-spectrogram reconstruction between input $\mathbf{X}$ and resynthesised output $\hat{\mathbf{X}}$:

$$\mathcal{L}_{recon}(G) = \sum_x ||\phi(\mathbf{X}) - \phi(\hat{\mathbf{X}})||_1, \tag{1.48}$$

where ϕ is a function computing Mel-spectrogram. The $\mathcal{L}_{fm}$ term is the distance between the discriminator activations of input x and resynthesised output $\hat{y}$:

$$\mathcal{L}_{fm}(D_j, G) = \sum_x \sum_{i=1}^{R} \frac{1}{M_i} ||\psi_i(\mathbf{X}) - \psi_i(\hat{\mathbf{X}})||_1, \tag{1.49}$$

where ψ_i is the function that extracts the activations of the i-th discriminator layer, and, M_i and R are the number of features in i and the number of layers in D_j, respectively.

While the concatenation of emotion label embeddings $\mathbf{z}_a$ to the input of generator G already conditions the resynthesised output $\hat{\mathbf{X}}$, we also included an SER-based loss to the loss function. To achieve this, we use a pre-trained SSL-based SER system E_{SER} introduced in [37]. The E_{SER} network was built by fine-tuning the `wav2vec2`-large-robust network [178] on the MSP-Podcast (v1.7) dataset [280]. The SER loss term $\mathcal{L}_{SER}$ is formulated as,

$$\mathcal{L}_{SER} = \sum_x [1 - \mathcal{L}_{CCC}(\mathbf{a}, E_{SER}(\hat{\mathbf{X}}))], \tag{1.50}$$

where $\mathcal{L}_{CCC}$ is the CCC loss [281] that measures similarity between two variables, $\mathbf{a}$ is the ground-truth emotion of input speech $\mathbf{X}$, and $E_{SER}(\hat{\mathbf{X}})$ is the predicted emotion for resynthesised speech $\hat{\mathbf{X}}$. The $\mathcal{L}_{CCC}$ measure varies between -1 and $+1$, where $+1$ denotes perfect similarity, therefore $1 - \mathcal{L}_{CCC}$ is minimized during training.

The final loss for generator G and discriminator D is:

$$\begin{aligned}\mathcal{L}_G(D, G) = &\sum_{j=1}^{J} [\mathcal{L}_{adv}(D_j, G) + \lambda_{fm}\mathcal{L}_{fm}(D_j, G)] \\ &+ \lambda_r \mathcal{L}_{recon}(G) + \lambda_{SER}\mathcal{L}_{SER},\end{aligned} \tag{1.51}$$

$$\mathcal{L}_D(D, G) = \sum_{j=1}^{J} \mathcal{L}_D(D_j, G), \tag{1.52}$$

where J is the number of sub-discriminators in D. Following [270], we set $\lambda_{fm} = 2$ and $\lambda_r = 45$. λ_{SER} was set to 1 after preliminary results supported this setup.

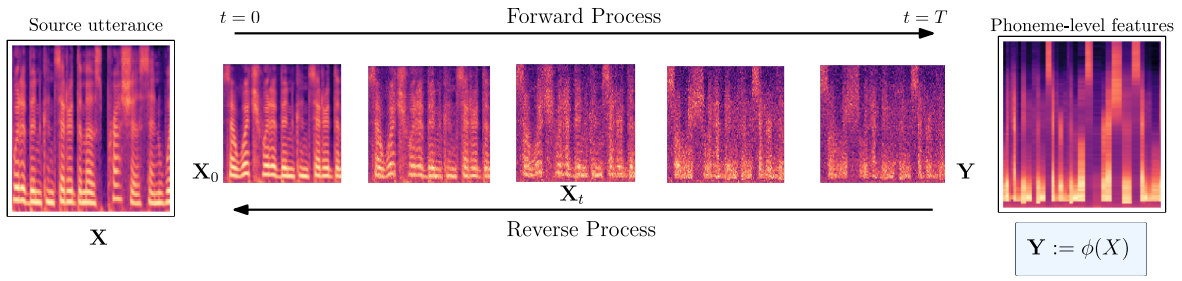


Figure 1.5: Forward–reverse diffusion processes: depiction of forward stochastic differential equation (SDE) (1.53) and reverse SDE (1.57) used in the training framework.

GANs as implicit generative models A fundamental limitation of GANs arises from their implicit formulation: although they are commonly referred to as "generative adversarial networks," they differ from *probabilistic* generative models in the sense defined by Murphy [48]. The generator implements a deterministic mapping from latent noise to the data space, without defining or optimizing an explicit likelihood function [60], [282]. In other words, the probabilistic formulation of emotional speech synthesis given in Eq. (1.45) is effectively replaced by learning a deterministic mapping from $\mathbf{z}$ to the data space of $\mathbf{X}$. The absence of a probabilistic, likelihood-based optimization of the data distribution has several consequences. First, from the standpoint of model training, likelihood evaluation provides a quantitative measure of how well a model explains observed data. Models with tractable likelihoods can be compared rigorously using metrics such as log-likelihood or the ELBO, whereas GANs cannot be evaluated in this principled manner. Second, likelihood maximization inherently promotes global coverage of the data distribution, thereby reducing the risk of *mode collapse*—a well-documented issue in adversarial training where the generator fails to capture the diversity of the target distribution [283], [284]. This limitation is particularly problematic for emotional speech synthesis: insufficient distribution coverage can lead to poor modeling of rare or extreme emotions, which are often omitted by the generator in a GAN-based framework. These shortcomings are also reflected in our experimental results, where the GAN-based SEC fail to well model extreme and less frequent emotional expressions, as discussed in detail in Section 4.1.

To overcome these issues and enhance the expressiveness of the SEC technique, we move towards the so-called *diffusion models*, which approach the task in a fully probabilistic and generative manner, thereby aligning more closely with the formulation in Eq. (1.45). While diffusion models do not allow for exact likelihood-based training or evaluation, they achieve the probabilistic modeling objective by learning a vector field that guides samples toward regions of high data probability, effectively capturing the structure of the underlying data distribution.

Diffusion based generation [P10]

Diffusion models are powerful generative models used across domains for tasks such as image editing [285], speech enhancement [273], [276], and text-to-speech (TTS) [64]. The central idea is to define a *forward* noising mechanism that progressively perturbs data into a simple, tractable prior, together with a learned *reverse* mechanism that inverts this corruption to synthesize new samples.

Forward process. In discrete time, the forward (Markov) diffusion uses a noise schedule

$\{\beta_t\}_{t=1}^{T}$:

$$q(\mathbf{X}_t \mid \mathbf{X}_{t-1}) = \mathcal{N}\big(\sqrt{1-\beta_t}\,\mathbf{X}_{t-1},\, \beta_t\mathbf{I}\big), \quad \mathbf{X}_0 \sim p_{\text{data}},$$

which yields the closed form $\mathbf{X}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{X}_0 + \sqrt{1-\bar{\alpha}_t}\,\varepsilon,\ \varepsilon \sim \mathcal{N}(0,\mathbf{I}),\ \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. In continuous time, the same idea is expressed as a *forward SDE* that transports p_{data} to a *terminal* distribution that is easy to sample at inference (typically $\mathcal{N}(0,\mathbf{I})$).

Reverse process. Under mild regularity assumptions, the forward SDE admits a well-defined reverse-time SDE [286]. The reverse process—stochastic or deterministic—maps the terminal distribution back to the data distribution by progressively removing noise. In practice, a neural network parameterizes either the score $\mathbf{s}_\theta(\mathbf{X},t) \approx \nabla_x \log p_t(\mathbf{X})$ (continuous-time view) or the noise term $\varepsilon_\theta(\mathbf{X}_t,t)$ (discrete-time view), with optional conditioning on side information (e.g., class labels or text prompts) during both training and sampling.

In the task of SEC with non-parallel, in-the-wild training data, we employ a diffusion model as a decoder operating on the disentangled representations $\mathbf{z}$. This replaces the HiFiGAN-based vocoder in the second stage of generation, under the hypothesis that diffusion models yield higher-quality speech and offer stronger conditioning on speech attributes. While this advantage has been demonstrated extensively in image synthesis [287], most diffusion architectures and training practices have been developed for two-dimensional inputs. A similar pattern appears in speech synthesis, where Mel-spectrograms are typically treated as images and used directly as inputs to diffusion models [64], [274], [277]. Although latent diffusion models (LDM) [62] provide a route to extend diffusion modeling to the *audio* domain more broadly, within *speech* the dominant practice remains to apply diffusion either to Mel-spectrogram [64] or short-time Fourier transform (STFT)-based representations [273], [276]. To address this, we modify the lexical representation used in the first stage of SSL-based disentanglement—originally paired with a HiFiGAN vocoder. Instead of discrete HuBERT-based lexical units, we adopt the speaker- and affect-independent "average voice" phoneme-level Mel features from [64] as the lexical representation $\mathbf{z}_l$. These features, denoted $\mathbf{Y}$ and illustrated in Figure 1.5, produce a robotic sounding speech when vocoded: intelligible lexical content is preserved, but speaker identity and affective cues are largely removed. Formally, let $\mathbf{Y} \coloneqq \phi(\mathbf{X})$ denote the "average voice" representation of source speech, where $\phi(\cdot)$ is a pretrained phoneme encoder. For processing, we employ a transformer-based encoder following [277], as used in voice conversion, with outputs dimension-matched to the source Mel-spectrogram $\mathbf{X}$.

Given these representations, the diffusion-based decoder follows the SDE formalism of [64]. Let $t \in [0,1]$ be the continuous diffusion time. The forward SDE is

$$\mathrm{d}\mathbf{X}_t = \frac{1}{2}\beta_t(\mathbf{Y}-\mathbf{X}_t)\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\mathbf{w}, \tag{1.53}$$

where $\mathbf{w}$ is a standard Wiener process [288], $\mathbf{X}_t$ is the process state with initial condition $\mathbf{X}$, and $\beta_t \geq 0$ is the noise schedule. Given $\mathbf{X}_0$, any state $\mathbf{X}_t$ follows a Gaussian [288, Section 5], referred to as the *perturbation kernel*:

$$p_{0t}(\mathbf{X}_t|\mathbf{X}_0,\mathbf{Y}) = \mathcal{N}_{\mathbb{C}}\Big(\mathbf{X}_t; \boldsymbol{\mu}(\mathbf{X}_0,\mathbf{Y},t), \sigma(t)^2\mathbf{I}\Big). \tag{1.54}$$

Within this process state $\mathbf{X}$ formulation, [64] show that the mean evolves as

$$\mu(\mathbf{X}_0,\mathbf{Y},t) = \alpha_t\mathbf{X}_0 + (1-\alpha_t)\mathbf{Y}, \tag{1.55}$$

with $\alpha_t = \exp(-\frac{1}{2}\int_0^t \beta_s \, \mathrm{d}s)$, and variance

$$\sigma(t)^2 = (1 - \alpha_t^2)\mathbf{I}. \tag{1.56}$$

We parameterize the schedule as $\beta_t = b_0 + t(b_1 - b_0)$ and choose $b_0, b_1 > 0$ such that $\alpha_1 \approx 0$. In this setting, the mean path interpolates from the source distribution at $t = 0$ toward the "average voice" distribution at $t = 1$. The forward SDE (1.53) admits a reverse-time SDE [286]:

$$\mathrm{d}\mathbf{X}_t = \left[-\frac{1}{2}\beta_t(\mathbf{Y} - \mathbf{X}_t) + \beta_t \nabla_{\mathbf{X}_t} \log p_t(\mathbf{X}_t|\mathbf{Y})\right] \mathrm{d}t + \beta_t \mathrm{d}\widetilde{\mathbf{w}}, \tag{1.57}$$

where $\mathrm{d}\widetilde{\mathbf{w}}$ is a Wiener process evolving backward in diffusion time. The reverse process approximately traces the forward trajectory in reverse: it starts near the "average voice" distribution and, as $t \to 0$, approaches the source-target distribution. This forward and reverse processes described above are depicted in Fig. 1.5.

We train a *score model* $\mathbf{s}_\theta(\mathbf{X}_t, \mathbf{Y}, \mathbf{z}_s, \mathbf{z}_a, t)$—instantiated as the U-Net of [277]—to approximate the score function $\nabla_{\mathbf{X}_t} \log p_t(\mathbf{X}_t|\mathbf{Y})$. With the trained $\mathbf{s}_\theta$, the reverse SDE generates an estimate of $\mathbf{X}_0$ from the "average voice" input $\mathbf{Y}$ conditioned on speaker identity $\mathbf{z}_s$ and affect embeddings $\mathbf{z}_a$. Intuitively, the diffusion-based decoder learns to reconstruct $\mathbf{X}_0$ while disentangling lexical (l), speaker (s), and affective (e) attributes—thereby removing the need for parallel data during training. During *inference*, a target affect embedding $E(\bar{e})$ is employed to convert the emotion of the source utterance to the given target affect. The target emotion embedding $E(\bar{e})$ is defined as the *averaged* affect embedding of a set reference utterance samples belonging to the affect category $\bar{e}$, as

$$E(\bar{e}) \coloneqq \frac{1}{|A_p(\bar{e})|} \sum_{\mathbf{X}_0 \in A_p(\bar{e})} E(\mathbf{X}_0), \tag{1.58}$$

where the set of reference samples $A_p(\bar{e})$ is defined to be the top $p = 20\%$ samples belonging to the particular target arousal $\bar{e}$.

The score model is trained on the *score matching* loss [61] which aims to approximate the score function. The score matching loss for $\mathbf{X}_0$ at time t is formulated as

$$\mathcal{L}_s(\mathbf{X}_t) = \mathbb{E}_{\epsilon_t}\left[||\mathbf{s}_\theta(X_t, t) + \sigma(t)^{-1}\epsilon_t||_2^2\right] \tag{1.59}$$

where $\mathbf{X}_t = \mu(t) + \sigma(t)\epsilon_t$ and ϵ_t is sampled from $\mathcal{N}(0, \sigma(t))$. In addition to $\mathcal{L}_s$, we follow [262] to use a mel spectrogram reconstruction loss for better conditioning on emotion attributes. $\mathcal{L}_m$ measures the L_1-norm:

$$\mathcal{L}_m(\hat{\mathbf{X}}_0) = \sum_x ||\mathbf{X}_0 - \hat{\mathbf{X}}_0||_1, \tag{1.60}$$

where $\hat{\mathbf{X}}_0$ is the mel spectrogram of synthesized speech. Note here that during the training of the score model it is expensive to obtain $\hat{\mathbf{X}}_0$ which requires solving the full reverse SDE. For this, in contrast to [262], we utilize a single-step approximation of $\hat{\mathbf{X}}_0$ by only relying on $\mathbf{X}_t$, $\mathbf{s}_\theta$, and $\mathbf{Y}$, which are available during training. We use Tweedie's formula [289] to approximate $\hat{\mathbf{X}}_0$ as

$$\hat{\mathbf{X}}_0 = \frac{\hat{\mu}(t) - (1 - \alpha_t)\,\mathbf{Y}}{\alpha_t}. \tag{1.61}$$

where $\hat{\mu}(t)$ is an estimate of $\mu(t)$ (1.55), and is formulated as $\hat{\mu}(t) = \mathbf{X}_t - (\mathbf{s}_\theta(X_t, t) * \sigma(t)^2)$. With that, the final loss function is

$$\mathcal{L}(\mathbf{X}_t, \hat{\mathbf{X}}_0) = \mathcal{L}_s(\mathbf{X}_t) + \lambda_t \mathcal{L}_m(\hat{\mathbf{X}}_0), \tag{1.62}$$

where λ_t is a weighting function depending on the current diffusion time-step t. Considering that $\mathbf{X}_t$ contains more Gaussian noise for larger t, we set $\lambda_t = 1 - t^2$, thereby weighting the reconstruction loss term ($\mathcal{L}_m(\hat{\mathbf{X}}_0)$) more for smaller t values and gradually decreasing the weights for larger t.

Duration controllable generation [P11]

The generative modeling approaches discussed above—both HiFiGAN- and diffusion-based—have pioneered SEC in in-the-wild, non-parallel settings by enabling controlled modification of local acoustic properties such as fundamental frequency, spectral envelope, and energy. However, they often *lack explicit mechanisms for controlling sound duration.* This limitation is particularly critical for modeling emotional expression in speech, where prosodic variation and timing play a central role. To address this gap, we introduce an additional duration modeling module that jointly learns emotional content and its corresponding temporal patterns during training. This enables explicit control over speech duration, allowing modification of speaking rate and rhythm to better reflect target emotions—all without relying on parallel data.

We base our duration modeling on the discrete HuBERT speech units $\mathbf{z}_l$, which represent the lexical content of the input speech $\mathbf{X}$. These speech tokens are discrete representations obtained via K-means clustering, and prior work has shown that the resulting clusters—and thus the units—are correlated with phoneme classes [271]. Importantly, the number of consecutive repetitions of a given unit naturally reflects the lexical duration of the corresponding phoneme. To that, we formulate duration modeling as follows: for $\mathbf{z}_l$ of input speech, we train a *duration predictor (DP)* to predict the consecutive repetition of discrete speech units $\mathbf{d}$, conditioned on emotion and speaker representations. These repetitions represent the durations of each lexical unit. Firstly, the frame-level $\mathbf{z}_l$ is de-duplicated to extract unit-level speech units, where the repetitions are ignored to obtain consecutive unique speech units. Secondly, this unit-level representation is fed as input to the predictor DP. To further achieve speaker and emotion conditioned duration modeling, we concatenate the speaker and emotion embeddings $(\mathbf{z}_s, \mathbf{z}_e)$ and pass them as an additional input to the predictor.

The predictor is a simple deterministic neural network comprising two convolution layers and a linear layer to predict $\mathbf{d}_i$ for respective unit-level speech units, where i is the index in the de-duplicated sequence of discrete speech units. As an example, if the speech units in $\mathbf{z}_l$ are $[1, 1, 2, 2, 2, 1, 3, 3, 3, 3]$, the de-duplicated sequence would be $[1, 2, 1, 3]$, and the target $\mathbf{d}$ would be $[2, 3, 1, 4]$. For stable training and to better account for outliers in durations, we predict durations in the log-scale: $\log(\mathbf{d})$, as suggested in [64]. During training, the predicted log-scale durations/repetitions are directly used in the loss function and the true frame-level $\mathbf{z}_l$ is used as the input to the HiFiGAN decoder. However, during inference, the predicted log durations $\widehat{\log(\mathbf{d})}$ are reversed back into duration units as follows:

$$\widehat{\mathbf{d}} = \min\left(1, e^{\widehat{\log(\mathbf{d})}+1}\right). \tag{1.63}$$

Finally, the reversed durations $\widehat{\mathbf{d}}$ are used to duplicate the unit-level speech units to obtain the duration modeled discrete lexical units $\widehat{\mathbf{z}}_l$. Note that, as per the resynthesis paradigm,

we use the estimated $\widehat{\mathbf{z}}_l$ only during inference, and during training the true $\mathbf{z}_l$ is used. The final input to the HiFiGAN decoder is the combined concatenated representation: $(\mathbf{z}_l, \mathbf{z}_s, \mathbf{z}_e)$ during training and $(\widehat{\mathbf{z}}_l, \mathbf{z}_s, \mathbf{z}_e)$ at inference time. The DP can then be trained using a simple loss function $\mathcal{L}_{dur}$ by employing a discriminative loss term such as mean squared error (MSE) or L_1.

Thesis Contributions 3

In this research direction of emotion-conditioned speech synthesis, our contributions are twofold: (1) We pioneer research in SEC and ESS tasks trained and evaluated on complex in-the-wild emotion datasets, moving beyond the simpler, acted, and scripted emotional speech commonly used in prior work. (2) We propose HiFiGAN- and diffusion-based architectures that eliminate the need for parallel emotional speech samples through a resynthesis-based training strategy. As illustrated in Figure 1.1, the combination of the HiFiGAN- and diffusion-based approaches along with the duration modeling, represents the the *response and synthesis aspect of social intelligence* in group interactions and is framed in this thesis as **RQ5**, **RQ6**, and **RQ7** in Section 1.7. The corresponding publications are [P9]–[P11].

1.7 Outline and Contributions

The overarching aim of this thesis is to develop systems endowed with *social intelligence*, with a particular emphasis on *affect* as the key social signal and *speech* as the primary behavioral cue modality. Within this overarching goal, the thesis addresses two main research directions: (1) the recognition of emotional expressions, and (2) the synthesis of emotional expressions. The structure of this dissertation is organized as follows.

Chapter 3: Recognition of Emotional Expressions This first chapter addresses the research direction of *recognition* of emotional expressions. With respect to the multilevel nature of emotional expressions, we address research challenges at both the individual-level and the group-level. In Section 3.1, we present contributions to *individual-level* emotion recognition, with particular emphasis on modeling label uncertainty. Subsequently, Section 3.2 focuses on *group-level* affect, detailing the collection of group-level annotations and the associated modeling approach.

Section 3.1: Individual-level Emotion Recognition

Research Questions

RQ1 How can we best model label uncertainty for individual-level emotion recognition, and what empirical gains—beyond improved characterization of uncertainty in emotion annotations—does this modeling yield?

RQ2 How can label uncertainty be modeled to robustly account for sparse and limited annotations, beyond the common assumption of Gaussian-distributed annotations?

Addressing **RQ1**, in Section 3.1, we move beyond single-point estimates in affect recognition and propose stochastic estimates that explicitly capture uncertainty in emotion labels. To this end, we adopt a probabilistic modeling approach, using BNNs, that provides a principled framework for handling label uncertainty, in contrast to deterministic multi-task learning methods that treat it in a simplified manner [P2]. In our formulation, emotion annotations are modeled as samples from a distribution, approximated with a Gaussian assumption, and the model is trained using a label-distribution learning loss function that captures variability across annotations. Empirical evaluations demonstrate a key contribution of this approach: it better represents the uncertainty inherent in emotion annotations (i.e., deviations from consensus averages due to subjectivity and ambiguity in affect perception). However, the Gaussian assumption introduces a trade-off: while the model captures label uncertainty more effectively, its estimates of the consensus-based average affect labels are less accurate [P1].

To address this trade-off, and in line with **RQ2**, we extend the formulation in Section 3.1 by modeling emotion annotations with a Student's t-distribution [P3]. Since public affective datasets typically include only a limited number of annotations per sample, the Gaussian assumption can be overly restrictive. The Student's t-distribution, by contrast, naturally accounts for the small-sample size by incorporating the number of annotations into the distributional modeling. Based on this formulation, we derive a corresponding Kullback–Leibler divergence loss and use it to train an estimator of the label distribution. Using both qualitative and quantitative analysis on a toy dataset, we demonstrate the advantages of this t-distribution approach over the Gaussian, particularly when the

number of annotations is small (e.g., 3–6 per sample). More importantly, we show that this formulation not only better captures label uncertainty but also improves the modeling of consensus-based affect labels. These improvements are especially pronounced in more complex datasets and under cross-corpus analyses, underscoring the robustness of probabilistic approaches to label uncertainty [P1].

Section 3.2: Group-level Affect Recognition

Research Questions

RQ3 In what ways can group-level affect annotations be advanced beyond existing approaches to achieve stronger alignment between theories in organizational psychology and methods in affect recognition?

RQ4 To what extent can group-level constructs, such as collective affect, be modeled in a data-driven framework that accommodates arbitrary group sizes and captures the inherent dynamics of affect?

The limited body of work on group-level affect methodologies and datasets has largely overlooked several critical aspects of this social construct: (i) its dynamic nature, which ebbs and flows across consecutive interaction segments (Eq. 1.2), (ii) its multidimensional structure (Eq. (1.3)), and (iii) the distinction between purposive and non-purposive groups, with most prior studies focusing on the latter. Collectively, these omissions highlight a significant *theory–method misalignment* between models of group affect in organizational behavior research and recognition methodologies developed in computer science. We address this gap in **RQ3** [P4]. To capture the *temporal dynamics* of group affect, we conducted pilot studies to iteratively refine the annotation window size, allowing for a more faithful representation of affective fluctuations. To address the complexity of annotating affect in *purposive groups*, we extensively trained organizational psychology students to generate context-aware, high-quality labels, with training materials incorporating video markers identified during the pilot phase. Together, these methodological advances align our annotation process more closely with organizational psychology theory, which conceptualizes group affect as a dynamic, multidimensional phenomenon emerging within purposive groups.

Addressing **RQ4**, Section 3.2, we propose a graph neural network (GNN) based data-driven architecture that advances beyond traditional statistical group-level features, such as aggregated measures of dyadic synchrony and mimicry [P4]. Grounded in social network theory, the GNN models group interactions as a graph, where individuals are represented as nodes and their interpersonal relationships as edges. Through an *attention*-based aggregation mechanism, the model can flexibly handle arbitrary group sizes, while an LSTM layer applied over group-level features across consecutive interaction segments captures the temporal dynamics of affect. Our experiments demonstrate that this GNN-based methodology outperforms synchrony- and mimicry-based handcrafted features in modeling group affect, highlighting the added value of temporal modeling and the integration of multimodal cues.

Chapter 4: Synthesis of Emotional Expressions

This chapter focuses on the second research direction, namely the *synthesis* of emotional expressions. Building on the conceptualization of affect introduced in Section 1.4.2 and the

datasets discussed in Section 1.4.3, the primary objective and contribution of this thesis within this domain is to develop robust emotion-conditioned speech synthesis methodologies applied to in-the-wild affective datasets.

Section 4.1, 4.2, and 4.3: Speech Emotion Conversion using Generative Models

Research Questions

RQ5 How can generative modeling approaches be applied to the task of speech emotion conversion in the challenging context of in-the-wild affect datasets?

RQ6 To what extent can diffusion models, given their probabilistic and generative modeling advantages over neural vocoders, enhance the capabilities of emotion-conditioned speech synthesis?

RQ7 How does the integration of a dedicated duration modeling module affect the performance of speech emotion conversion, and can such a module be trained on in-the-wild data without relying on ground-truth duration information from parallel speech corpora?

Addressing **RQ5**, in Section 4.1, we introduce a HiFiGAN-based neural vocoder for speech emotion conversion (SEC) [P5]. A central challenge in working with in-the-wild datasets is the absence of parallel samples, which necessitates unsupervised training strategies. To enable emotion-conditioned generation during inference, our approach first disentangles input speech into lexical, speaker, and emotion attributes using SSL-based pretrained encoders, and then reconstructs speech through a neural vocoder conditioned on these factors. At inference, the emotion attribute is modified to synthesize speech in the target emotion. Results show that this framework, trained without parallel data, can generate natural-sounding emotion-conditioned speech, with stronger performance at mid-range emotion levels than at extremes.

Building on this, **RQ6** in Section 4.2 investigates diffusion models as a more powerful alternative to neural vocoders for SEC [P10]. We propose *EmoConv-Diff*, a GradTTS-based diffusion model [64] trained with a reconstruction loss that uses a single-step approximation of synthesized speech to reduce computational cost. By leveraging the probabilistic and generative strengths of diffusion, *EmoConv-Diff* achieves significant gains over the vocoder baseline, particularly in synthesizing speech at extreme emotional intensities—thereby addressing a key limitation observed in RQ5.

Finally, a common limitation of both vocoder- and diffusion-based models—which are typically non-autoregressive—is the absence of *explicit duration control*, as duration is often not directly modeled in such architectures. Addressing this in **RQ7**, Section 4.3, we propose a duration modeling framework based on resynthesis-driven discrete content representations [P11]. This approach learns to predict repetitions of lexical-based units, enabling modification of speech duration to match target emotions without requiring parallel data. We demonstrate that the inclusion of duration modeling substantially enhances emotional expressiveness: low-arousal emotions are reflected in longer durations and slower rates, while high-arousal emotions yield shorter, faster speech.

2

Overview of the Related Publications

The list of publications in this thesis follows the research plan outlined in Section 1.7. Peer-reviewed papers included in the body of this cumulative dissertation are highlighted in a blue box, while all other related publications and pre-prints are presented in Appendix A. Related abstract presentations are presented in Appendix B.

Chapter 3. Recognition of Emotional Expressions

3.1. Individual-level Emotion: Addressing Label Uncertainty Modeling

[P1] N. Raj Prabhu, N. Lehmann-Willenbrock, and T. Gerkmann, "End-to-end label uncertainty modeling in speech emotion recognition using bayesian neural networks and label distribution learning," *IEEE Transactions on Affective Computing*, vol. 15, no. 2, pp. 579–592, 2024.

[P2] N. Raj Prabhu, G. Carbajal, N. Lehmann-Willenbrock, and T. Gerkmann, "End-to-end label uncertainty modeling for speech-based arousal recognition using Bayesian neural networks," in *Proceedings of Interspeech*, Incheon, Korea, Sep. 2022.

[P3] N. Raj Prabhu, N. Lehmann-Willenbrock, and T. Gerkmann, "Label uncertainty modeling and prediction for speech emotion recognition using t-distributions," in *IEEE International Conference on Affective Computing and Intelligent Interaction*, Nara, Japan, Oct. 2022.

3.2. Group-level Affect: Annotations and Multimodal Modeling

[P4] N. Raj Prabhu, M. Tsfasman, C. Oertel, T. Gerkmann, and N. Lehmann-Willenbrock, "Dynamics of collective group affect: Group-level annotations and the multimodal modeling of convergence and divergence," *Accepted for IEEE Transactions on Affective Computing*, Dec. 2025.

A. Other Related Publication

[P5] D. de Oliveira, N. Raj Prabhu, and T. Gerkmann, "Leveraging semantic information for efficient self-supervised emotion recognition with audio-textual distilled models," in *Proceedings of Interspeech*, Dublin, Ireland, 2023.

[P6] N. Raj Prabhu, V. Begemann, N. Lehmann-Willenbrock, and T. Gerkmann, "Ground-truth of affect labels: Accounting for subjectivity, ambiguity, and uncertainty," in *Computational group and team dynamics: Forging an interdisciplinary science*, S. Kozlowski, H. Hung, N. Lehmann-Willenbrock, and A. Salah, Eds. United Kingdom: Oxford University Press, 2025, In Press.

B. Abstract Presentations

[P7] V. Begemann, C. S. Hemshorn de Sanchez, N. Raj Prabhu, T. Gerkmann, and N. Lehmann-Willenbrock, "Starting on the same note: How pre-discussion small talk and paraverbal synchrony contribute to group entitativity," in *19th Annual INGRoup Conference*, Jul. 2024.

[P8] V. Begemann, C. S. Hemshorn de Sanchez, N. Raj Prabhu, and N. Lehmann-Willenbrock, "In sync with sustainability: Acoustic-prosodic synchrony and attitudes toward the climate crisis in group discussions," in *53rd Deutsche Gesellschaft für Psychologie (DGP) Congress*, Sep. 2024.

Chapter 4. Synthesis of Emotional Expressions

4.1. Speech Emotion Conversion using Neural Vocoder

[P9] N. Raj Prabhu, N. Lehmann-Willenbrock, and T. Gerkmann, "In-the-wild speech emotion conversion using disentangled self-supervised representations and neural vocoder-based resynthesis," in *Proceedings of ITG Conference Speech Communication*, Sep. 2023.

4.2. Speech Emotion Conversion using Diffusion Models

[P10] N. Raj Prabhu, B. Lay, S. Welker, N. Lehmann-Willenbrock, and T. Gerkmann, "EMOCONV-Diff: Diffusion-based speech emotion conversion for non-parallel and in-the-wild data," in *Proceedings of IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, Apr. 2024, pp. 11 651–11 655.

4.3. Duration Modeling for Speech Emotion Conversion

[P11] N. Raj Prabhu, D. de Oliveira, N. Lehmann-Willenbrock, and T. Gerkmann, "Enhancing in-the-wild speech emotion conversion with resynthesis-based duration modeling," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Dec. 2025.

3

Recognition of Emotional Expressions

3.1 Individual-level Emotion: Addressing Label Uncertainty Modeling [P1]

Abstract

To train machine learning algorithms to predict emotional expressions in terms of arousal and valence, annotated datasets are needed. However, as different people perceive others' emotional expressions differently, their annotations are subjective. To account for this, annotations are typically collected from multiple annotators and averaged to obtain ground-truth labels. However, when exclusively trained on this averaged ground-truth, the model is agnostic to the inherent subjectivity in emotional expressions. In this work, we therefore propose an end-to-end Bayesian neural network capable of being trained on a distribution of annotations to also capture the subjectivity-based label uncertainty. Instead of a Gaussian, we model the annotation distribution using Student's t-distribution, which also accounts for the number of annotations available. We derive the corresponding Kullback-Leibler divergence loss and use it to train an estimator for the annotation distribution, from which the mean and uncertainty can be inferred. We validate the proposed method using two in-the-wild datasets. We show that the proposed t-distribution based approach achieves state-of-the-art uncertainty modeling results in speech emotion recognition, and also consistent results in cross-corpora evaluations. Furthermore, analyses reveal that the advantage of a t-distribution over a Gaussian grows with increasing inter-annotator correlation and a decreasing number of annotations available.

Reference

N. Raj Prabhu and N. Lehmann-Willenbrock and T. Gerkmann, "End-to-End Label Uncertainty Modeling in Speech Emotion Recognition Using Bayesian Neural Networks and Label Distribution Learning", in *IEEE Transactions on Affective Computing*, vol. 15, no. 2, pp. 579-592, April-June 2024. DOI: 10.1109/TAFFC.2023.3283595.

Copyright Notice

Authors' Contributions

Navin Raj Prabhu led the study, including the initial conceptualization, algorithm development, neural network training, experimental validation, and manuscript preparation. Nale Lehmann-Willenbrock contributed by reviewing the manuscript and helping to refine the argumentation and overall framing. Timo Gerkmann provided key insights into the experimental validation, offered valuable methodological feedback through discussions, and participated in the manuscript review.

End-to-End Label Uncertainty Modeling in Speech Emotion Recognition using Bayesian Neural Networks and Label Distribution Learning

Navin Raj Prabhu, *Student Member, IEEE,* Nale Lehmann-Willenbrock, *Non-Member, IEEE,* and Timo Gerkmann, *Senior Member, IEEE,*

Abstract—To train machine learning algorithms to predict emotional expressions in terms of arousal and valence, annotated datasets are needed. However, as different people perceive others' emotional expressions differently, their annotations are subjective. To account for this, annotations are typically collected from multiple annotators and averaged to obtain ground-truth labels. However, when exclusively trained on this averaged ground-truth, the model is agnostic to the inherent subjectivity in emotional expressions. In this work, we therefore propose an end-to-end Bayesian neural network capable of being trained on a distribution of annotations to also capture the subjectivity-based label uncertainty. Instead of a Gaussian, we model the annotation distribution using Student's t-distribution, which also accounts for the number of annotations available. We derive the corresponding Kullback-Leibler divergence loss and use it to train an estimator for the annotation distribution, from which the mean and uncertainty can be inferred. We validate the proposed method using two in-the-wild datasets. We show that the proposed t-distribution based approach achieves state-of-the-art uncertainty modeling results in speech emotion recognition, and also consistent results in cross-corpora evaluations. Furthermore, analyses reveal that the advantage of a t-distribution over a Gaussian grows with increasing inter-annotator correlation and a decreasing number of annotations available.

Index Terms—Emotional expressions, annotations, Bayesian neural networks, label distribution learning, end-to-end, speech emotion recognition, uncertainty, subjectivity, t-distributions, Kullback-Leibler divergence loss

1 Introduction

Emotions are typically studied as emotional expressions that others subjectively perceive and respond to [1], [2]. A standard theoretical backdrop for analyzing emotions is the two-dimensional pleasure and arousal framework [3], which describes emotional expressions along two continuous, bipolar, and orthogonal dimensions: pleasure-displeasure (*valence*) and activation-deactivation (*arousal*). One way emotions become expressed in social interactions, and therefore accessible for social signal processing (SSP), concerns speech signals. Speech emotion recognition (SER) research spans roughly two decades [2], with ever improving state-of-the-art techniques. As a consequence, research on SER has shown increasing prominence in highly-critical and socially relevant domains, such as health, security, and employee well-being [2], [4], [5].

A crucial challenge when studying emotional expressions in terms of arousal and valence is that their annotations are per se *subjective* because different people perceive others' emotional expressions differently [2], [5]. To address this, these annotations are typically collected by multiple annotators, and consensus on ground-truth is reached using techniques such as average scores [6], majority voting [6], or evaluator-weighted mean (EWE) [7]. These techniques in principle can lead to loss of valuable information on the inherently subjective nature of emotional expressions, and also tend to mask less prominent emotional traits [5]. In the context of *reliability* in real-world applications, SER systems not only need to model ground-truth labels but also account for the subjectivity inherent in these labels [2], [8]. Moreover, by also capturing subjectivity, SER systems can be efficiently deployed in human-in-the-loop solutions, and aid in the development of algorithms for active learning, co-training, and curriculum learning [5].

In this work, we tackle the problem of recognising emotional expressions using speech signals, in terms of *time*- and *value*-continuous arousal and valence. To this, we adopt an *end-to-end* learning framework. Common SER approaches rely on hand-crafted features to model emotion labels [9], [10]. Recently, end-to-end architectures have been shown to deliver state-of-the-art emotion predictions [11]–[13], by *learning* features rather than relying on hand-crafted features. For modeling *subjectivity* in emotions, scholars have suggested that end-to-end learning also promotes learning subjectivity dependent representations [14].

Uncertainty in machine learning (ML) is generally investigated in terms of two broader categories. First, *model uncertainty*, or epistemic uncertainty, accounts for the uncertainty in model parameters, and the resulting uncertainty *can* be reduced given enough data-samples [15]–[17]. Second, *label uncertainty*, or aleatoric uncertainty, captures noise inherent in the data-samples, such as sensor noise or label noise [15],

- *Navin Raj Prabhu and Timo Gerkman are with the Signal Processing Lab, Universität Hamburg, Germany, 20146. E-mail: navin.raj.prabhu@uni-hamburg.de, timo.gerkmann@uni-hamburg.de*
- *Nale Lehmann-Willenbrock is with the Department of Industrial and Organizational Psychology, Universität Hamburg, Germany, 20146. E-mail: nale.lehmann-willenbrock@uni-hamburg.de*

This work was supported by the Landesforschungsförderung Hamburg (LFF-FV79), as part of the research unit "Mechanisms of Change in Dynamic Social Interactions".

[16]. Label uncertainty *cannot* be reduced even if more data-samples are collected. Label uncertainty has been further categorized into *homoscedastic* uncertainty, which remains constant across data-samples, and *heteroscedastic* uncertainty, whose uncertainty depends on the respective data-sample. This work specifically aims to model the heteroscedastic label uncertainty, henceforth simply mentioned as *label uncertainty*, that corresponds to the *inherent subjectivity in emotion annotations*.

We propose to use *Bayes by Backpropagation* (BBB), a Bayesian neural network (BNN) technique, in order to capture label uncertainty. In ML, stochastic and probabilistic models have mainly been used for uncertainty modeling, through ensemble learning [18], encoder-decoder architectures [19], neural processes [20], [21], and BNNs [22]–[24]. Among these, the Bayesian frameworks show improved performance over non-Bayesian baselines in previous works [15], [25], making BNNs such as Monte-Carlo dropout [22] and BBB [23] promising candidates for modeling label uncertainty in SER. BBB learns a distribution over weights to produce *stochastic outputs*, which makes it capable of being trained on a distribution of annotations.

With BBB capable of being trained on a distribution of annotations, we capture label uncertainty using the *label distribution learning* (LDL) technique [25], leveraging Kullback-Leibler (KL) divergence-based loss functions. Subjective annotations of emotion create a label distribution to represent the subjectivity in emotions [5]. For simplicity, histograms [5], [26], and Gaussians [27] have been employed to represent the label distributions. However, Gaussians and histograms with *limited* and *sparse* observations are sensitive to outliers thereby losing their robustness in this scenario [28]–[30]. Note here that publicly available SER datasets commonly comprise only limited annotations (e.g., 3 to 6) [31]–[35], and there is consensus in the literature that gaining more annotations is expensive and resource inefficient [5], [36]. At the same time, a significant degree of subjectivity in annotations is also well noted [10], [14], thereby leading to sparse annotations with outliers. To tackle this, in this work, we model emotion annotation distributions as a Student's t-distribution, or simply t-distribution. Kotz et al. [29], and, Bishop [28], note that in scenarios of limited and sparse observations with outliers, the t-distribution becomes more robust over a Gaussian, by producing robust mean and standard deviation estimates of the distribution.

We derive a KL divergence loss for label uncertainty that quantifies distribution similarity between stochastic emotion predictions, modeled as a Gaussian distribution, and *ground-truth emotion annotations*, modeled as a *t-distribution*. Subsequently, we present analyses to reveal the benefits of using t-distribution over a Gaussian. We validate the proposed model in two in-the-wild datasets, AVEC'16 [37] and MSP-Conversation [31]. We show that the proposed model can aptly capture label uncertainty with state-of-the-art results for both datasets, along with a robust loss curve. To emphasize the benefits of the t-distribution, we present experiments studying the impact of the number of emotion annotations available. Finally, we perform an ablation study to understand specific benefits of the respective modules in the architecture.

This work is based on two prior conference contributions [38], [39], which to the best of our knowledge are the first in the literature to use BBB and LDL in SER. These works were also the first to tackle the problem of limited emotion annotations from an ML perspective. Previously, we only validated the method in one dataset, and with limited experiments [38], [39]. In this extension, we additionally validate the method in a larger and more complex dataset, the MSP-Conversation [31], along with cross-corpora evaluations. This extension is also the first in the literature to present SER results in this novel dataset [31]. Existing analyses and experiments from [38], [39] were also extended to MSP-Conversation. Moreover, we performed additional experiments that include an experiment to understand the impact of the number of annotations available, and an ablation study. Code for the proposed model and loss function is available online [1].

2 Background and Related Work

2.1 Ground-truth labels

To handle subjectivity in emotional expressions, annotations $\{y_1, y_2, .., y_a\}$ for emotions are typically collected from multiple annotators (a) [33], [35]. The *ground-truth label* is then obtained as the mean m across all annotations from a annotators [40],

$$m = \frac{1}{a}\sum_{i=1}^{a} y_i. \tag{1}$$

Alternatively, the EWE, which weights annotations with inter-annotator correlations, has been proposed as the *gold-standard* $\widetilde{m}$ [7]. Both m and $\widetilde{m}$ based approximation of ground-truth leads to loss of information on subjectivity [5].

Given a raw audio sequence of T frames $\mathcal{X} = [x_1, x_2, ..., x_T]$, traditional SER approaches aim to estimate either the m_t or $\widetilde{m}_t$ for each time frame $t \in [1, T]$, referred to as $\widehat{m}_t$. The concordance correlation coefficient (CCC) [41] has been widely used as a loss function for this task [2]. For Pearson correlation r, the CCC between m and $\widehat{m}$, for T frames is:

$$\mathcal{L}_{\text{CCC}}(m) = \frac{2r\sigma_m\sigma_{\widehat{m}}}{\sigma_m^2 + \sigma_{\widehat{m}}^2 + (\mu_m - \mu_{\widehat{m}})^2}, \tag{2}$$

where $\mu_m = \frac{1}{T}\sum_{t=1}^{T} m_t$, $\sigma_m^2 = \frac{1}{T}\sum_{t=1}^{T}(m_t - \mu_m)^2$, and $\mu_{\widehat{m}}$, $\sigma_{\widehat{m}}^2$ are obtained similarly for $\widehat{m}$. The CCC metric measures the agreement between two variables, in our case the ground-truth m and its estimate $\widehat{m}$. It ranges from -1 to $+1$, with perfect agreement at $+1$. In contrast to Pearson's correlation r, CCC takes both the linear correlation and the bias in to account, which makes it preferable over Pearson's correlation as the loss and evaluation metric in SER.

2.2 Label uncertainty in SER

As an alternative to exclusively modeling m_t or $\widetilde{m}_t$, previous research has attempted to model ground truth that also explains inter-annotator disagreement, for example by means of soft labels [5] and entropy of disagreement [42]. Sridhar et al. [5] proposed an auto-encoder technique that jointly models soft- and hard-labels of emotion annotations

1. https://github.com/sp-uhh/label-uncertainty-ser

and subsequently estimates label uncertainty as the entropy on soft-labels. Fayek et al. [43] and Tarantino et al. [44] proposed to learn soft labels instead of m_t with improved performance. Steidl et al. [42] quantified label uncertainty using the entropy measure and trained a model to minimize the difference in entropy between model outputs and annotator disagreement.

Label uncertainty has also been approached as a prediction task by estimating the moments of a distribution [9], [45]. Han et al. [9], [45] used a multi-task learning (MTL) framework to model the unbiased standard deviation s of a annotators as an auxiliary task,

$$s = \sqrt{\frac{1}{a-1}\sum_{i=1}^{a}(y_i - m)^2}. \tag{3}$$

Similarly, Dang et al. [46] captured the temporal dependencies in the annotation signals, using multi-rater Gaussian mixture regression and Kalman filters. Sridhar et al. [10] introduced a Monte-Carlo (MC) dropout model to obtain uncertainty estimates from the distribution of stochastic outputs. However, their model was not explicitly trained on any label uncertainty estimate and hence could only capture the model uncertainty, but not the label uncertainty. A similar MC dropout was used by Rizos et al. [47], who proposed a meta-learning framework that uses uncertainty estimates to potentially detect highly-uncertain samples and perform soft data selection for the training process.

Research efforts have also been made to estimate emotion annotations as a *distribution*, using LDL [26], [27], [38], [39]. Foteinopoulou et al. [27] trained an MTL network using a KL divergence loss that models emotion annotations as a *uni-variate Gaussian* with mean m and unknown variance. Chou et al. [26] used LDL to convert subjective annotations into *histogram*-based distributional labels for training. In our preliminary work [38], we modeled emotion annotations as a *Gaussian* using BBB-based uncertainty modeling. Notwithstanding the improved performance of these approaches, a drawback concerns the limited annotations on which previous *histogram* or a *Gaussian* assumptions were based [26], [27], [38]. These assumptions are susceptible to unreliable m and s for lower values of a and sparsely distributed annotations [28], [29]. In our subsequent work [39] and in this extension, we tackle this problem by modeling emotion annotation distribution as a t-distribution and show advantages over a Gaussian assumption.

2.3 On distributions

A Gaussian distribution $\mathcal{Y} \sim \mathcal{N}(\mu, \sigma^2)$ is a continuous probability distribution for a real-valued random variable y, with the general form of its probability density function [28]:

$$p(y \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}. \tag{4}$$

The parameters μ and σ are the mean and standard deviation of the distribution, respectively. Due to its simplicity, Gaussians are often used to model random variables whose distributional family are unknown [23], [38]. However, Gaussians are sensitive to outliers, especially in cases of *limited* and *sparse* observations of the random variable

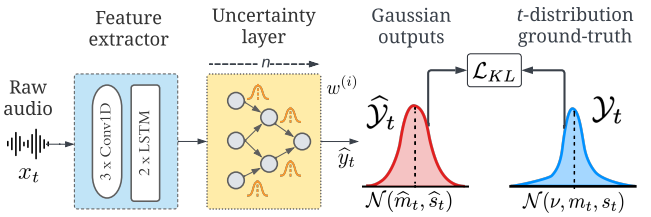


Fig. 1: Overview of proposed architecture and loss $\mathcal{L}_{\text{KL}}$. n: number of forward passes. $w^{(i)}$ and $\widehat{y}_t$: stochastically sampled weight and realization of $\widehat{\mathcal{Y}}_t$, at i^{th} forward pass.

[28]. In this case, the t-distribution is noted to become more robust and realistic over a Gaussian [28], [29].

Student's t-distribution, $\mathcal{Y}_t \sim \mathcal{N}(\nu, \mu, \sigma)$, arises when estimating the moments of a normally distributed population in *situations where the sample size is small* [29], [48], with the probability density function given by [30], [49],

$$p(y \mid \nu, \mu, \sigma) = \frac{1}{\mathrm{B}(\frac{1}{2}, \frac{\nu}{2})} \frac{1}{\sqrt{\nu\sigma^2}} \left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}, \tag{5}$$

where ν denotes the degrees of freedom and B(., .) is the Beta function, for Gamma function Γ, formulated as,

$$B(i, j) = \frac{\Gamma(i)\,\Gamma(j)}{\Gamma(i+j)}. \tag{6}$$

The density function (5) is symmetric, and its overall shape resembles the bell shape of a normally distributed variable, except that it has heavier tails, meaning that it better captures values that fall far from its mean (i.e., outliers) [28], [29]. The degree of freedom ν, also known as the normality parameter, controls the normality of the distribution and is correlated with the standard deviation of the distribution σ [28], [29]. In (5), the standard deviation σ takes the scaled form, where σ is scaled using the normality parameter ν:

$$\sigma\sqrt{\frac{\nu}{\nu-2}} \text{ for } \nu > 2, \tag{7}$$

As ν increases, the t-distribution approaches the normal distribution [30]. The normality parameter ν, in our case, allows the t-distribution to also account for the number of annotations available.

The robustness of the t-distribution, in cases of *limited* and *sparse* observations of the random variable, is associated with its ability to better capture the outliers by also accounting for the number of observations of the random variable [28]. This is the key motivation behind using the t-distribution to model the emotion annotations, to produce robust mean and standard deviation estimates by also accounting for the number of annotations available.

3 Proposed Label Uncertainty Model

In order to better represent subjectivity in emotional expressions, we estimate the *emotion annotation distribution* $\mathcal{Y}_t$ for each time-frame t, given raw audio x_t. While the true distributional family of subjectively perceived emotions $\mathcal{Y}_t$ is unknown, for simplicity, we can assume that it follows a Gaussian distribution:

$$\mathcal{Y}_t \sim \mathcal{N}(m_t, {s_t}^2). \tag{8}$$

However, with only a limited number of annotations, and in cases where the annotations are sparsely distributed with outliers, we argue that a Gaussian assumption is rather crude [28], [29]. Instead, we propose to model the emotion annotations as a t-distribution, with degrees of freedom ν:

$$\mathcal{Y}_t \sim \mathcal{N}(\nu, m_t, {s_t}^2). \tag{9}$$

Thus, the goal is to obtain an estimate $\widehat{\mathcal{Y}}_t$ of $\mathcal{Y}_t$ and infer both $\widehat{m}_t$ and $\widehat{s}_t$ from realizations of $\widehat{\mathcal{Y}}_t$.

3.1 End-to-end DNN architecture

We propose an end-to-end architecture that uses a feature extractor to learn temporal-paralinguistic features from x_t, and an uncertainty layer to estimate $\mathcal{Y}_t$ (see Fig. 1). The feature extractor, inspired by [11], consists of three Conv1D layers followed by two stacked long short term memory (LSTM) layers. The uncertainty layer is devised using the BBB technique [23], comprising three BBB-based MLP.

3.2 Model uncertainty loss

Unlike a standard neuron which optimizes a deterministic weight w, the BBB-based neuron learns a probability distribution on the weight w by calculating the variational posterior $P(w|\mathcal{D})$ given the training data $\mathcal{D}$ [23]. Intuitively, this regularizes w to also capture the inherent uncertainty in $\mathcal{D}$. In contrast to learning a deterministic weight w to exclusively estimate m_t, the BBB neuron learns a Gaussian weight distribution $\mathcal{N}(\mu_w, \sigma_w)$, thereby allowing the model to not only estimate m_t but also s_t. Estimation of s_t is achieved by calculating the standard deviation of the stochastic estimates obtained from stochastically sampled weights $w^{(i)}$. To obtain a non-negative estimate of the standard deviation of the weight distribution σ_w, we re-parameterize the standard deviation as $\sigma_w = \log(1 + \exp(\rho_w))$ based on an initial estimate ρ_w. This way $\theta = (\mu_w, \rho_w)$ can be optimized using simple backpropagation and still ensure a non-negative σ_w.

For an optimized θ, the predictive distribution $\widehat{\mathcal{Y}}_t$ for x_t, is given by $P(\widehat{y}_t|x_t) = \mathbb{E}_{P(w|D)}[P(\widehat{y}_t|x_t, w)]$, where $\widehat{y}_t$ are realizations of $\widehat{\mathcal{Y}}_t$. Unfortunately, the expectation under the posterior of weights is intractable. To tackle this, [23] proposed to learn θ of a weight distribution $q(w|\theta)$, the variational posterior, that minimizes the Kullback-Leibler (KL) divergence with the true Bayesian posterior, resulting in the negative evidence lower bound (ELBO),

$$f(w, \theta)_{\text{BBB}} = \text{KL}\big[q(w|\theta)\|P(w)\big] - \mathbb{E}_{q(w|\theta)}\big[\log P(D|w)\big]. \tag{10}$$

In BBB, stochastic outputs are achieved using multiple forward passes n with stochastically sampled weights w, thereby modeling $\widehat{\mathcal{Y}}_t$ using the n stochastic estimates. To account for the stochastic outputs, (10) is approximated as,

$$\mathcal{L}_{\text{BBB}} \approx \sum_{i=1}^{n} \log q(w^{(i)}|\theta) - \log P(w^{(i)}) - \log P(D|w^{(i)}). \tag{11}$$

where $w^{(i)}$ denotes the i^{th} weight drawn from $q(w|\theta)$. The BBB window-size b controls how often new weights are sampled for time-continuous SER. The degree of uncertainty is assumed to be constant within this time period. During testing, the uncertainty estimate $\widehat{s}_t$ is the standard deviation of $\widehat{\mathcal{Y}}_t$, and, $\widehat{m}_t$ is the realization $\widehat{y}_t$ obtained using the mean of the optimized weights μ_w. Obtaining $\widehat{m}_t$ using μ_w helps overcome the randomization effect of sampling from $q(w|\theta)$, which showed better performances in our case.

Note that variables n, a, and ν are closely related to one another. The three variables all denote the number of samples used to model distribution, either $\widehat{\mathcal{Y}}_t$ or $\mathcal{Y}_t$. Variable n represents the number of forward passes, thereby the number of stochastic estimates used to model the *estimate* distribution $\widehat{\mathcal{Y}}_t$. Variable a represents the number of annotations used to model the ground-truth distribution $\mathcal{Y}_t$. In the probability density function of a t-distribution (5), ν denotes the degree of freedom. In this work, the ν of a t-distribution is set to a enabling the *ground-truth* distribution $\mathcal{Y}_t$ to also account for the number of annotations available.

3.3 Label uncertainty loss

While (11) exclusively captures *model uncertainty*, the aim of this work is to also capture *label uncertainty*. For this, using LDL, inspired by [16], we introduce a *KL divergence-based loss* to fit our model to the annotation distribution $\mathcal{Y}_t$, with either a Gaussian assumption (in Sec. 3.3.1) or a t-distribution assumption (in Sec. 3.3.2).

3.3.1 Gaussian $\mathcal{Y}_t$ KL divergence

For a Gaussian assumption on $\mathcal{Y}_t$ (8), the label uncertainty loss, the KL divergence between two Gaussians $\mathcal{Y}_t \sim \mathcal{N}(m_t, {s_t}^2)$ and $\widehat{\mathcal{Y}}_t \sim \mathcal{N}(\widehat{m_t}, {\widehat{s_t}}^2)$ is formulated as [28],

$$\mathcal{L}_{KL}(\mathcal{Y}_t\|\widehat{\mathcal{Y}}_t) = \log\left(\frac{\widehat{s_t}}{s_t}\right) + \frac{{s_t}^2 + (m_t - \widehat{m_t})^2}{2{\widehat{s_t}}^2} - \frac{1}{2}. \tag{12}$$

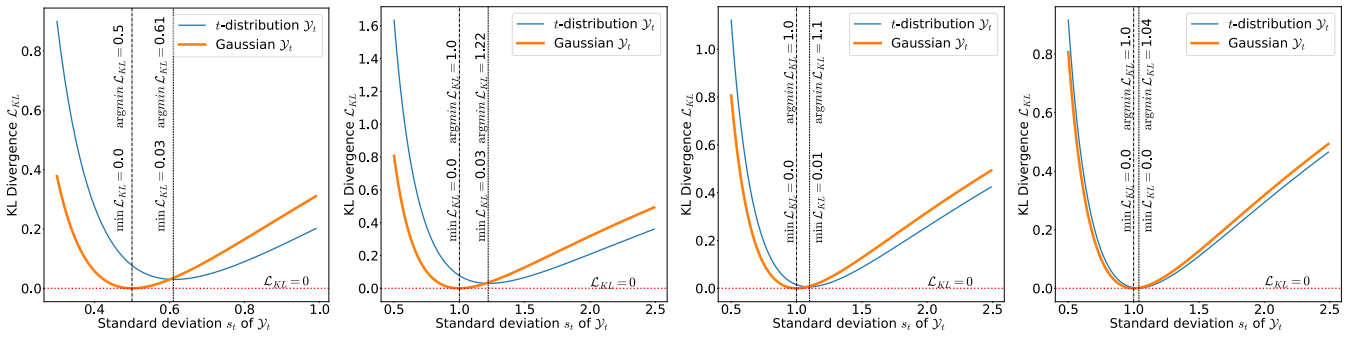
The KL divergence is asymmetric, making the order of distributions crucial. In (12), we choose $\widehat{\mathcal{Y}}_t$ to follow $\mathcal{Y}_t$, for a mean-seeking approximation, rather than a mode-seeking one, to capture the full distribution [50, p. 76]. See Supplementary Sec. 3 for further details on the choice between mean- and mode-seeking approximation using $\mathcal{L}_{\text{KL}}$.

3.3.2 t-distribution $\mathcal{Y}_t$ KL divergence

For $\mathcal{Y}_t$ as a t-distribution (9), we derive the KL divergence between $\mathcal{Y}_t \sim \mathcal{N}(\nu, m_t, {s_t}^2)$ and the Gaussian outputs $\widehat{\mathcal{Y}}_t \sim \mathcal{N}(\widehat{m_t}, {\widehat{s_t}}^2)$. Assuming a Gaussian on $\widehat{\mathcal{Y}}$ is fair, as the number of stochastic outputs to model $\widehat{\mathcal{Y}}$ can be controlled using n in (11). In this work, we intend to fix $n \geq 30$, as a t-distribution converges to a stable Gaussian with 30 samples [30], [49]. As a positive side effect, we result in deriving the KL divergence between a Gaussian and a t-distribution, in contrast to between two t-distributions, with the latter involving mathematical complexities in calculating intractable expectations for a loss function.

For a Gaussian $\widehat{\mathcal{Y}}$ (see (4)), and a t-distributed $\mathcal{Y}$ (see (5)), the $\mathcal{L}_{KL}$ is formulated as [51], [52],

$$\mathcal{L}_{KL}(\mathcal{Y}_t\|\widehat{\mathcal{Y}}_t) = H(\mathcal{Y}_t, \widehat{\mathcal{Y}}_t) - H(\mathcal{Y}_t), \tag{13}$$



(a) $\widehat{\mathcal{Y}}_t \sim \mathcal{N}(0, 0.5)$, and $\nu = 6$ (b) $\widehat{\mathcal{Y}}_t \sim \mathcal{N}(0, 1)$, and $\nu = 6$ (c) $\widehat{\mathcal{Y}}_t \sim \mathcal{N}(0, 1)$, and $\nu = 12$ (d) $\widehat{\mathcal{Y}}_t \sim \mathcal{N}(0, 1)$, and $\nu = 30$

Fig. 2: Analysis of the t-distribution based KL divergence $\mathcal{L}_{KL}$ (16), in comparison with Gaussian $\mathcal{L}_{KL}$ (12).

where $\mathrm{H}(\cdot,\cdot)$ is the cross-entropy between two distributions, and $\mathrm{H}(\cdot)$ is the entropy of a distribution. The cross-entropy term $\mathrm{H}(\cdot,\cdot)$ in (13), using (4), can be further formulated as,

$$\begin{aligned}
\mathrm{H}(\mathcal{Y}_t, \widehat{\mathcal{Y}}_t) &= -\int \mathcal{Y}_t(y)\ \log\ \widehat{\mathcal{Y}}_t(y)\, dy \\
&= \frac{1}{2}\log(2\pi\widehat{s_t}^2) + \int \mathcal{Y}_t(y)\ \Big(\frac{(y-\widehat{m_t})^2}{2\widehat{s_t}^2}\Big) dy \\
&= \frac{1}{2}\log(2\pi\widehat{s_t}^2) + \frac{1}{2\widehat{s_t}^2}\Big[\int \mathcal{Y}_t(y) y^2\, dy \\
&\qquad -2\widehat{m_t}\int \mathcal{Y}_t(y)\, y\, dy + \widehat{m_t}^2 \int \mathcal{Y}_t(y)\, dy\Big]. \quad (14)
\end{aligned}$$

Noting that $\int \mathcal{Y}_t(y) y^2\, dy = {m_t}^2 + {s_t}^2$, $\int \mathcal{Y}_t(y)\, y\, dy = m_t$, and $\int \mathcal{Y}_t(y)\, dy = 1$, where m_t and s_t are parameters of the t-distribution $\mathcal{Y}_t$, $p(y \mid \nu, m_t, s_t)$, the equation (14) becomes,

$$\begin{aligned}
&= \frac{1}{2}\log(2\pi\widehat{s_t}^2) + \frac{1}{2\widehat{s_t}^2}\Big[{s_t}^2 + {m_t}^2 - 2\widehat{m_t}m_t + \widehat{m_t}^2\Big] \\
&= \frac{1}{2}\log(2\pi\widehat{s_t}^2) + \frac{{s_t}^2 + (m_t - \widehat{m_t})^2}{2\widehat{s_t}^2} \quad (15)
\end{aligned}$$

Finally, using (15) in (13), our proposed KL divergence is

$$\boxed{\mathcal{L}_{KL} = \frac{1}{2}\log(2\pi\widehat{s_t}^2) + \frac{{s_t}^2 + (m_t - \widehat{m_t})^2}{2\widehat{s_t}^2} - H(\mathcal{Y}_t).} \quad (16)$$

We implement (16) as a custom loss function by extending the `studentT` pytorch sub-package [53].

3.3.3 *Comparing Gaussian and t-distribution loss*

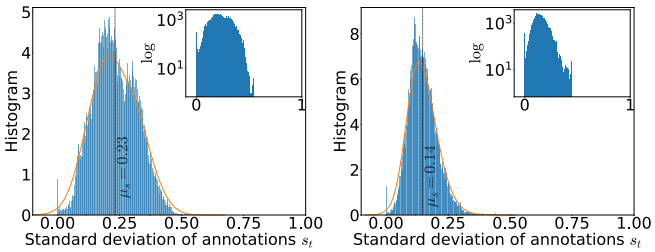
While the two loss-functions (12) and (16) have their second term in common, two differences can be noted. Firstly, as (12) calculates the divergence between two similar distributions, $\mathcal{Y}_t$ and $\widehat{\mathcal{Y}}_t$, it includes the logarithm of the ratio between the two Gaussian's standard deviation in its formulation. However, in (16), the deviations of $\mathcal{Y}_t$ and $\widehat{\mathcal{Y}}_t$ are *separately* quantified using terms $\frac{1}{2}\log(2\pi\widehat{s_t}^2)$ and $H(\mathcal{Y}_t)$, respectively. Secondly, (16) is dependent on the number of annotations available through scaling s_t with the normality factor ν (7).

To further understand the advantages of the t-distribution $\mathcal{L}_{KL}$ (16) over the Gaussian $\mathcal{L}_{KL}$ (12), we plot the $\mathcal{L}_{KL}$ values as a function of varying s_t, for (16) and (12). We perform this analysis under four different scenarios, for different values of $\widehat{s_t}$ and ν, i) Figure 2a for scenario $\widehat{s_t} = 0.5$ and $\nu = 6$, ii) Figure 2b for scenario $\widehat{s_t} = 1.0$ and $\nu = 6$, iii) Figure 2c for scenario $\widehat{s_t} = 1.0$ and $\nu = 12$, and, iv) Figure 2d for scenario $\widehat{s_t} = 1.0$ and $\nu = 30$.

From Figure 2, firstly, we see that $\mathcal{L}_{KL}$ behaves differently when the ground-truth $\mathcal{Y}_t$ is modeled as a t-distribution (16), in comparison to the Gaussian assumption (12). Specifically, from Figure 2a, for $\widehat{s_t} = 0.5$ and $\nu = 6$, we see that the minimum $\mathcal{L}_{KL}$ (16) is achieved only at $s_t = 0.61$, in contrast to the Gaussian (12) $\widehat{s_t} = s_t = 0.5$. While the Gaussian attempts exactly fitting the model to the ground-truth $s_t = 0.5$, the t-distribution tries to fit on a more relaxed $s_t = 0.61$ by also considering the reduced degree of freedom $\nu = 6$. This behaviour is similar to the confidence intervals calculation using a t-distribution [54, Sec. 9.5], where relaxation on s_t is noted with respect to ν. Moreover, [28] associate this relaxed s_t towards the increased robustness of the t-distribution to sparse distributions with outliers.

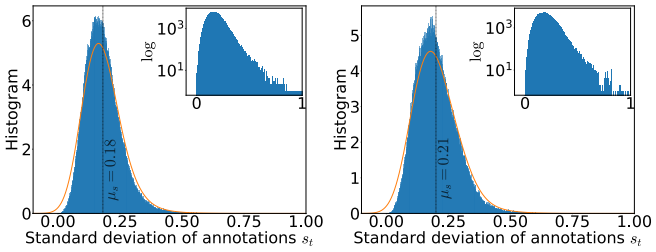
Secondly, we note that the observed relaxation on s_t is dependent on two factors, 1) the standard deviation of the stochastic outputs $\widehat{s_t}$, and 2) the degree of freedom of the ground-truth ν. From figures 2a and 2b, we see that, while ν is constant, the relaxation on s_t *increases* along with an increase in $\widehat{s_t}$. At $\widehat{s_t} = 0.5$ a relaxation of 0.11 is made by the t-distribution (16) from 0.5 to 0.61, while a larger relaxation of 0.22 is made for $\widehat{s_t} = 1.0$. Similarly, from figures 2c and 2d, we see that, while $\widehat{s_t}$ is constant, as ν increases the relaxation on s_t *decreases*. That is, the t-distribution (16) starts behaving similar to a Gaussian, in line with literature that states that as the degree of freedom ν of t-distribution increases, the distribution converges into a Gaussian [29], [30], [49]. This is also in line with our initial motivation behind using the t-distribution, which we expected to account for the number of annotations available while fitting on annotation distribution $\mathcal{Y}$.

From an ML and SER perspective, from Figure 2, we note several benefits of t-distribution based loss term towards label uncertainty modeling. Firstly, training on a t-distribution $\mathcal{L}_{KL}$ (16) leads to training on a relaxed s_t, and can lead to better capturing of the *whole* ground-truth label distribution. In other words, this can lead to the t-distribution better accounting for sparse annotations with outliers, where a relatively high likelihood is associated along the tails of the distribution, as noted by Bishop [28]. Secondly, we note that in all cases, the t-distribution $\mathcal{L}_{KL}$ (16) values are always



(a) For arousal. (b) For valence.

Fig. 3: Histogram of standard deviations s_t in AVEC'16.



(a) For arousal. (b) For valence.

Fig. 4: Histogram of standard deviations s_t in MSPConv.

higher than Gaussian $\mathcal{L}_{KL}$ for lower values of s_t and $\widehat{s_t}$. This might lead to larger penalization of the model through the $\mathcal{L}_{KL}$ loss, and may thereby promote better and quicker convergence during training, in comparison to the Gaussian $\mathcal{L}_{KL}$ (12). Finally, the t-distribution $\mathcal{L}_{KL}$ (16) can also adapt to different datasets by also accounting for the number of annotations available during training.

3.4 Training loss

The proposed end-to-end uncertainty loss is formulated as,

$$\mathcal{L} = (1 - \mathcal{L}_{\text{CCC}}(m)) + \mathcal{L}_{\text{BBB}} + \alpha \mathcal{L}_{\text{KL}}. \tag{17}$$

Intuitively, $\mathcal{L}_{\text{CCC}}(m)$ optimizes for mean predictions m, $\mathcal{L}_{\text{BBB}}$ optimizes for BBB weight distributions, and $\mathcal{L}_{\text{KL}}$ optimizes for the label distribution $\mathcal{Y}_t$. For $\alpha = 0$, the model only captures model uncertainty (MU). For $\alpha = 1$, the model also captures *label uncertainty* (MU+LU or t-LU). $\mathcal{L}_{\text{CCC}}(m)$ is used as part of $\mathcal{L}$ to achieve faster convergence and jointly optimize for mean predictions. Including $\mathcal{L}_{\text{CCC}}(m)$ might lead to better optimization of the feature extractor [11], [55].

In Equation (17), α is the tuning parameter that decides how much we want to regularize our model to also account for the label uncertainty. While the proposed models only use two values for the α (0 and 1), as an additional study, we also experimented with varying regularization on the label uncertainty loss term $\mathcal{L}_{\text{KL}}$ (see Supplementary Sec. 5).

4 Experimental Setup

4.1 Dataset

To validate our proposed methodology, we use two publicly available in-the-wild datasets, with time- and value-continuous annotations for arousal and valence. Firstly, the AVEC'16 [37] version of the RECOLA dataset [33], which has *2.15hrs* of annotated dyadic interactions. Secondly, the MSP-Conversation dataset, which has *15.15hrs* of annotated interactions with groups of 2-7 interlocutors.

4.1.1 AVEC'16 dataset

The dataset consists of arousal and valence annotations by $a = 6$ annotators at 40 ms frame-rate, or 25 frames per second (fps). The arousal and valence annotations in the dataset are distributed on average with $\mu_m = 0.01$ and $\mu_m = 0.11$, and $\mu_s = 0.23$ and $\mu_s = 0.14$, respectively, where $\mu_s = \frac{1}{T}\sum_{t=1}^{T} s_t$. Further, in Figure 3 the distribution of s_t is illustrated. It can be noted from Fig. 3 that s_t distributions are skewed towards high standard deviations s_t, thereby indicating the high-level of subjectivity present in the dataset. The high skewness is even more evident in the log-histogram plotted along in Fig. 3. The dataset is divided into speaker disjoint partitions for training, development, and testing, with nine 300 s recordings each. Results with respect to the AVEC'16 are presented only in terms of the development partition, as the annotations for the test partition are not publicly available. Similarly, the hyperparameters are fine-tuned on the train partition for this particular dataset. Note that the *posterior distribution* $P(w|D)$ and the time-shift for post-processing are the only parameters tuned using the partitions. See Supplementary Sec. 1 for the complete list of hyperparameters used.

4.1.2 MSP-Conversation dataset

The MSP-Conversation, or simply *MSPConv*, is approximately 7 times larger than AVEC'16, comprising of in-the-wild podcasts. The wide range of podcast recordings leads to high variability in terms of population size, group size, and more importantly its emotional content [31], [32], making the MSPConv a more complex dataset to model.

The dataset consists of time- and value-continuous annotations for arousal and valence, performed by at least $a = 6$ annotators at ≈ 16 ms frame-rate, or 60 fps, however not uniform in all cases [31]. For uniform sampling rate, we perform median filtering with a window-size of 500ms, as suggested in [31]. To keep the sampling rate consistent between the two datasets, for cross-corpora evaluations, we use a step-size of $1/25$s in median filtering. A local normalization, i.e., for each annotated sequence and for each annotator, was performed using zero-mean unit-deviation normalization [33], similar to AVEC'16. As illustrated in Figure 4, in the MSPConv dataset [31], arousal and valence annotations are distributed on average with $\mu_m = -0.01$ and $\mu_m = 0.00$, and $\mu_s = 0.18$ and $\mu_s = 0.21$, respectively. Further revealing the complexity of MSPConv, when comparing figures 3 and 4, we see that the level of subjectivity in MSPConv is higher than the AVEC'16 dataset, where in MSPConv the s_t distribution tail is more skewed towards high subjectivity. The high skewness is more evident in the log-histogram plotted along in Fig. 4.

In preliminary experiments, we noted that the arousal and valence annotations were prone to periodic distortion noises, especially from particular annotators— *001*, *007*, and *009*. This could have originated from any technical error or from a human error by the annotator. Directly training on these noisy annotations degraded the performance of

TABLE 1: Comparison on mean m, standard deviation s, and label distribution estimations $\mathcal{Y}$, in terms of $\mathcal{L}_{\text{ccc}}(m)$, $\mathcal{L}_{\text{ccc}}(s)$, and $\mathcal{L}_{\text{KL}}$, respectively. Larger CCC indicates improved performance as indicated by ↑. Lower KL indicates improved performance as indicated by ↓. ** indicates that the respective approach achieves statistically significant better results than *all* other approaches in comparison. * indicates that it achieves statistically significant better results over *only some* of the approaches in comparison. Results in brackets (.) are for the respective development partition of the dataset.

	Arousal			**Valence**		
	$\mathcal{L}_{\text{ccc}}(m)\uparrow$	$\mathcal{L}_{\text{ccc}}(s)\uparrow$	$\mathcal{L}_{\text{KL}}\downarrow$	$\mathcal{L}_{\text{ccc}}(m)\uparrow$	$\mathcal{L}_{\text{ccc}}(s)\uparrow$	$\mathcal{L}_{\text{KL}}\downarrow$
E2E Baseline w/o Temp	0.581	-	-	0.129	-	-
E2E Baseline [11]	0.770	-	-	0.361	-	-
STL [9]	0.727	-	-	0.389	-	-
MTL PU [9]	0.740	0.310	0.776	**0.420****	0.032	0.960
MU [38]	0.762	0.077	0.675	0.332	0.040	0.631
MU+LU [38]	0.751	0.361	0.250	0.301	0.048	0.405
t-LU (proposed)	**0.782****	**0.381****	**0.228****	0.400*	**0.050***	**0.386****

(a) Quantitative results on **AVEC'16** dataset.

	Arousal			**Valence**		
	$\mathcal{L}_{\text{ccc}}(m)\uparrow$	$\mathcal{L}_{\text{ccc}}(s)\uparrow$	$\mathcal{L}_{\text{KL}}\downarrow$	$\mathcal{L}_{\text{ccc}}(m)\uparrow$	$\mathcal{L}_{\text{ccc}}(s)\uparrow$	$\mathcal{L}_{\text{KL}}\downarrow$
E2E Baseline w/o Temp	0.177 (0.206)	-	-	0.080 (0.115)	-	-
E2E Baseline [11]	0.373 (0.407)	-	-	0.192 (0.183)	-	-
STL [9]	0.292 (0.360)	-	-	0.190 (0.189)	-	-
MTL PU [9]	0.296 (0.363)	0.107 (0.105)	0.527 (0.440)	0.181 (0.185)	0.030 (0.030)	0.560 (0.450)
MU [38]	0.367 (0.406)	0.052 (0.067)	0.380 (0.410)	0.208 (0.220)	0.022 (0.028)	0.451 (0.439)
MU+LU [38]	0.357 (0.397)	0.111 (0.123)	0.370 (0.322)	0.191 (0.219)	0.029 (0.032)	0.410 (0.396)
t-LU (proposed)	**0.389**** (0.421**)	**0.118*** (0.134*)	**0.357**** (0.317**)	**0.213*** (0.224*)	**0.032*** (0.035*)	**0.373**** (0.382**)

(b) Quantitative results on **MSPConv** dataset.

all models in comparison. Ignoring the noisy annotations might lead to a loss of information, and might also result in a reduced number of available annotations to derive ground-truth. To reduce these periodic distortions and still retain the inherent annotation information, we use a low-pass filter [56] with a cut-off frequency of 0.25Hz. The cut-off frequency was tuned using a Fourier analysis [57] followed by a qualitative analysis of the filtered annotations. Filtering was performed only on annotations *with periodic distortions*, i.e., from the three annotators– *001*, *007*, and *009*.

4.2 Baselines and Proposed model versions

E2E Baselines: This baseline is a reimplementation of [11], with the same end-to-end framework as our proposed model but a multi-layer perceptron instead of the uncertainty layer. The model does not capture any form of uncertainty, and is exclusively trained on the $\mathcal{L}_{\text{CCC}}(m)$ loss (2).

Time-continuous ground-truth annotations contain temporal dependencies [58], where an annotation at time t can be expected to have a high correlation with annotations at time $t+1$ and $t-1$. Our proposed architecture accounts for this temporal dependency using two stacked LSTM layers. Moreover, temporal modeling is achieved by batching annotations into sequences of 12s each (300 frames of 40ms each). With this setup, the LSTM operation is performed over the sequence rather than over a single frame, thereby directly learning temporal dependencies. To assess the impact of this temporal modeling, we use an additional baseline: *E2E Baseline w/o Temp* where the *LSTM operation is performed on the feature dimension*, in contrast to the E2E Baseline where the operation is performed on the temporal dimension. This way the number of parameters is kept the same for the two allowing for a fair comparison.

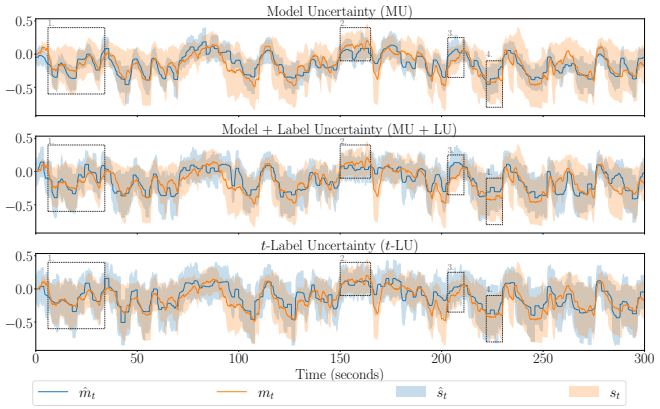
MTL Baselines: From [9], [45], as the baselines, we use the perception uncertainty (*MTL PU*) and single-task models (*STL*). The MTL PU is a label uncertainty model that also models s_t as an auxiliary task. The STL does not capture uncertainty and is exclusively trained on $\mathcal{L}_{\text{CCC}}(m)$ (2). For a fair comparison, we reimplemented these baselines. Crucially, the reimplementation also enables us to compare the models in terms of their standard deviation s estimates, which were not presented in Han et al.'s work [9].

Proposed BBB-LDL versions: We use three versions of the proposed label uncertainty model. Firstly, the *Model Uncertainty* (*MU*) version, which shares the same DNN architecture as the other BBB version but is trained on (17) with $\alpha = 0$. Secondly, the *Label Uncertainty* (*MU+LU*) version also captures the label uncertainty and is trained on (17) with $\alpha = 1$. The MU+LU version however makes a Gaussian assumption on $\mathcal{Y}_t$, thereby $\mathcal{L}_{\text{KL}}$ follows (12). Finally, the *t-distribution Label Uncertainty* (*t-LU*) version, which is trained on the same loss function (17) but models $\mathcal{Y}_t$ as a t-distribution, and $\mathcal{L}_{\text{KL}}$ follows (16).

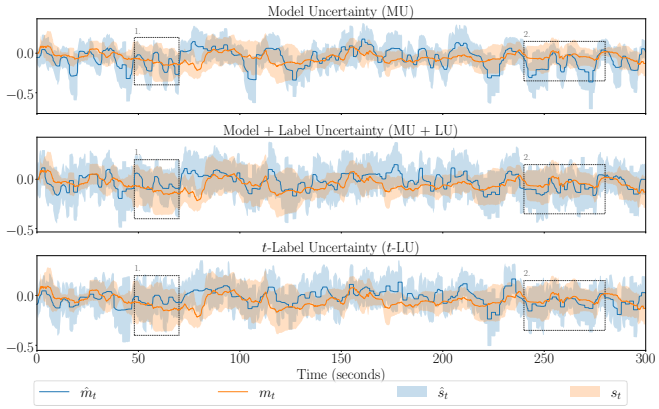
Finally, for all the models two post-processing techniques are applied, namely, median filtering [11] and time-shifting [59] (with shifts between 0.04s and 10s). See supplementary Sec. 2 for further detailed information.

4.3 Validation measures

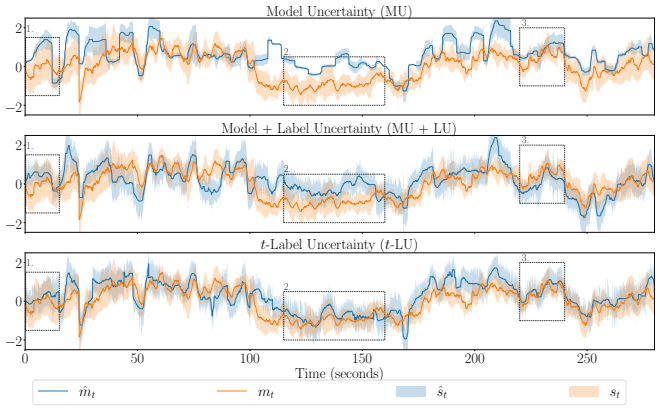
To validate the proposed method's *mean* and *standard deviation* estimates, we use $\mathcal{L}_{\text{CCC}}(m)$ and $\mathcal{L}_{\text{CCC}}(s)$ metrics, respectively, widely used in literature [9], [11], [55]. However, $\mathcal{L}_{\text{CCC}}(m)$ and $\mathcal{L}_{\text{CCC}}(s)$ validate mean and standard deviation estimates *separately*. To further *jointly* validate mean and standard deviation estimates, as label distribution $\widehat{\mathcal{Y}}_t$, we use the $\mathcal{L}_{\text{KL}}$ measure. For a fair comparison, we validate



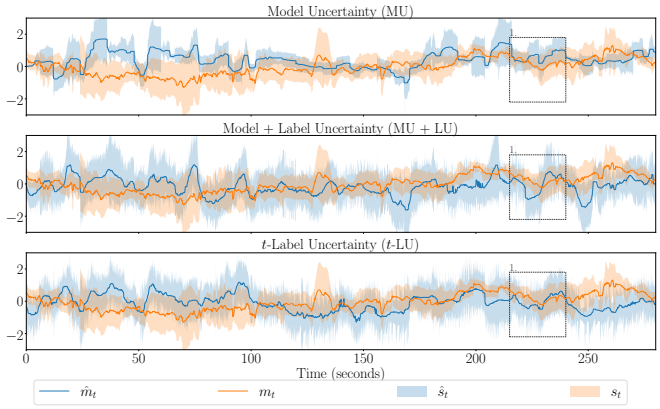
(a) For arousal, in **AVEC'16** dataset.



(b) For valence, in **AVEC'16** dataset.



(c) For arousal, in **MSPConv** dataset.



(d) For valence, in **MSPConv** dataset.

Fig. 5: Label distribution $\mathcal{Y}_t$ estimation results for a test subject.

all the models in comparison using $\mathcal{L}_{\text{KL}}$ based on their respective distribution assumptions on $\mathcal{Y}_t$, as the models are trained in a similar fashion. The proposed t-LU version is validated and trained on the t-distribution $\mathcal{L}_{\text{KL}}$ (16), and the baselines on the Gaussian $\mathcal{L}_{\text{KL}}$ (12). Nevertheless, from the experiments, we also noted that the proposed t-LU performs better in terms of both (16) and (12). Finally, the statistical significance of results is estimated using a one-tailed t-test, asserting significance for p-values ≤ 0.05.

5 Results and Discussion

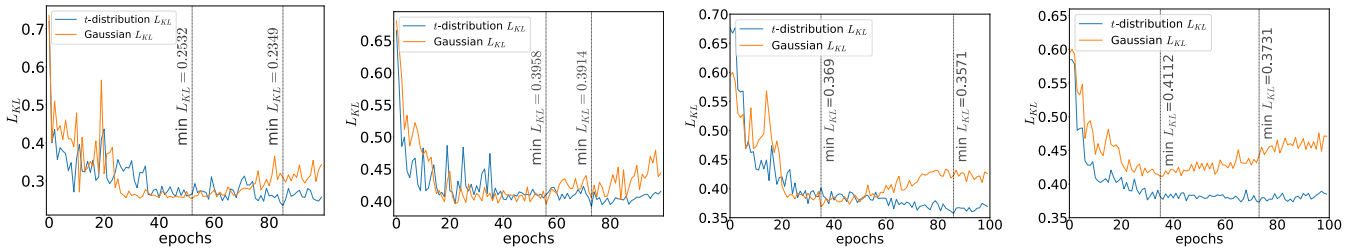
5.1 Quantitative analysis of estimates

Table 1 shows the average performance of the baselines and the proposed models, in terms of their mean m, standard deviation s, and distribution $\widehat{\mathcal{Y}}_t$ estimations, $\mathcal{L}_{\text{CCC}}(m)$, $\mathcal{L}_{\text{CCC}}(s)$, and $\mathcal{L}_{KL}$, respectively. Results are presented with respect to two datasets, for AVEC'16 in Table 1a, and for MSPConv in Table 1b. From the analysis presented in Section 4.1, we note that the MSPConv is a more complex dataset, in terms of modeling label uncertainty.

5.1.1 Comparison on mean estimates

In terms of *arousal*, Table 1 shows that the proposed t-LU model performs the *best* in comparison with the baselines, in both AVEC'16 (Table 1a) and MSPConv (Table 1b) datasets, with *statistical significance*. Four key takeaways can be noted from the $\mathcal{L}_{\text{CCC}}(m)$ results for *arousal*. *Firstly*, the proposed BBB-LDL versions (MU, MU+LU, and t-LU) achieve better $\mathcal{L}_{\text{CCC}}(m)$ than the MTL baselines (STL, and MTL PU). In the more challenging MSPConv dataset, the performance improvement is even more evident, which highlights the robustness of the proposed approach. For example, while the t-LU improves over MTL PU by 0.042 in AVEC'16, a larger improvement of 0.093 $\mathcal{L}_{\text{CCC}}(m)$ can be noted in the MSPConv. *Secondly*, between the BBB-LDL versions, the superiority of the proposed t-distribution $\mathcal{L}_{KL}$ (16) over the Gaussian $\mathcal{L}_{KL}$ (12) is noted, with t-LU outperforming MU+LU in both the datasets. *Thirdly*, when incorporating uncertainty modeling in the E2E Baseline, a *compromise* on $\mathcal{L}_{\text{CCC}}(m)$ is made with improving uncertainty estimates ($\mathcal{L}_{\text{CCC}}(s)$ and $\mathcal{L}_{\text{KL}}$). This can be noted when comparing the results of MU and MU+LU with that of the E2E Baseline. However, the proposed t-LU is free from this compromise, outperforming the E2E Baseline and other BBB-LDL versions. The t-LU achieves a $\mathcal{L}_{\text{CCC}}(m)$ of 0.782 in AVEC'16 and 0.389 in MSPConv, with E2E Baseline achieving 0.770 and 0.373, respectively. *Finally*, the *E2E Baseline w/o Temp* performs the worst in comparison. The improved performance of E2E Baseline and the proposed models over the *E2E Baseline w/o Temp* emphasises the fact that temporal modeling exists in the proposed models and is achieved through the inclusion of the LSTM layers in their architecture.

In terms of *valence*, in the AVEC'16, the MTL PU baseline



(a) For arousal, in **AVEC'16**. (b) For valence, in **AVEC'16**. (c) For arousal, in **MSPConv**. (d) For valence, in **MSPConv**.

Fig. 6: Loss curve comparison between Gaussian $\mathcal{L}_{KL}$ (12) and proposed t-distribution $\mathcal{L}_{KL}$ (16).

performs significantly better than the proposed models. However, in the larger and more complex MSPConv dataset, the proposed t-LU performs the best with statistical significance. The MTL PU requires dataset dependent tuning of the loss using the average correlation between m_t and s_t. For example, $\mathcal{L} = \mathcal{L}_{\text{ccc}}(m) - \mathcal{L}_{\text{ccc}}(s)$ for datasets with negative average correlation, and $\mathcal{L} = \mathcal{L}_{\text{ccc}}(m) + \mathcal{L}_{\text{ccc}}(s)$ for a positive one [9]. In AVEC'16 the average correlation between m_t and s_t is $+0.103$ (a positive correlation exists). In MSPConv the average correlation between m_t and s_t is 0.002 where no correlation exists. With these statistics, we note that the MTL PU is robust only in datasets where a correlation between m_t and s_t exists, and not robust in cases of complex datasets like the MSPConv. Moreover, with the dataset dependent tuning of the loss, MTL PU is also not robust in cross-corpora evaluation (see Sec. 5.4).

5.1.2 Comparison on uncertainty estimates

Table 1 shows that the proposed t-LU achieves the best uncertainty estimates across datasets, in terms of both $\mathcal{L}_{\text{CCC}}(s)$ and $\mathcal{L}_{KL}$. In AVEC'16, the improvements are *statistically significant* over all baselines in comparison. In MSPConv, the improvements are statistically significant over all baselines only with respect to the $\mathcal{L}_{KL}$ measure. In terms of the $\mathcal{L}_{\text{CCC}}(s)$ measure, improvements are not statistically significant over the MU+LU baseline alone. For instance, in AVEC'16, t-LU achieves 0.381 $\mathcal{L}_{\text{CCC}}(s)$ and 0.228 $\mathcal{L}_{KL}$, improving with statistical significance. In MSPConv, t-LU achieves 0.118 $\mathcal{L}_{\text{CCC}}(s)$ and 0.357 $\mathcal{L}_{KL}$, where statistical significance over *all* other baselines exists only for $\mathcal{L}_{KL}$. The reason for this trend is that, firstly, the MSPConv is more complex with larger levels of subjectivity (see Sec. 4.1). Secondly, the model is exclusively trained on $\mathcal{L}_{KL}$, so direct improvements over $\mathcal{L}_{KL}$ is expected rather than on $\mathcal{L}_{\text{CCC}}(s)$.

For *valence* in AVEC'16, unlike the $\mathcal{L}_{\text{CCC}}(m)$ performances, Table 1a shows that the proposed t-LU achieves improved *uncertainty estimates*, in terms of both the measures ($\mathcal{L}_{\text{CCC}}(s)$ and $\mathcal{L}_{KL}$). Moreover, the improvements are statistical significance over all other baselines in terms of the $\mathcal{L}_{KL}$ measure, but only over the MTL-based baselines in terms of the $\mathcal{L}_{\text{CCC}}(s)$ measure. Similar improvement trends can also be noted in the more complex MSPConv dataset (from Table 1b). This improved uncertainty estimates of the proposed t-LU across datasets emphasises the advantage of using the t-distribution based $\mathcal{L}_{KL}$ loss (16) for label uncertainty modeling. The t-distribution, as seen in Figure 2, promotes the model to fit on a more relaxed s_t, thereby more robust in capturing the whole label distribution. The fitting on a relaxed s_t leads to increased robustness towards outliers, as noted in [28].

5.2 Qualitative analysis of estimates

For qualitative analyses, we plot the mean $\widehat{m}_t$ and standard deviation $\widehat{s}_t$ estimates of $\widehat{\mathcal{Y}}_t$ against the m_t and standard deviation s_t of ground-truth distribution $\mathcal{Y}_t$. Plots for a test subject from AVEC'16, in terms of arousal and valence, can be seen in figures 5a and 5b, respectively, and, for MSPConv, in figures 5c and 5d, respectively. Parts of the plots are boxed and numbered to note clear performance differences.

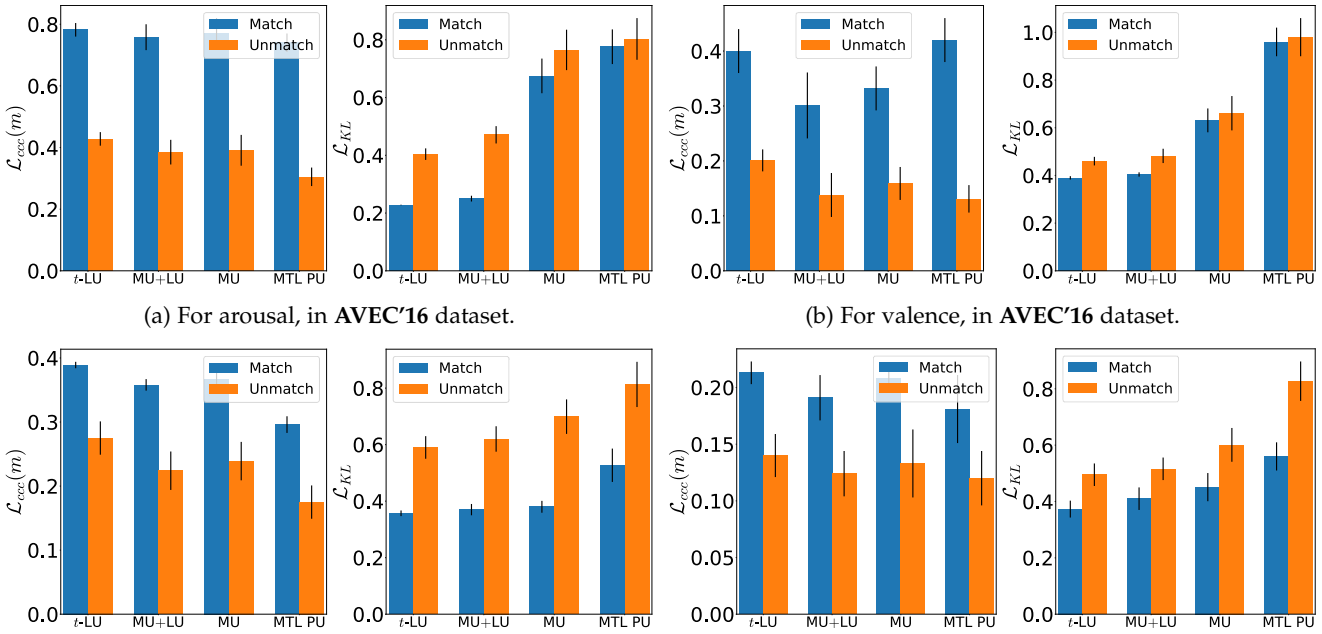
For *arousal*, in figures 5a and 5c, further backing the results in Table 1, the proposed t-LU model best captures m_t and s_t of the annotation distribution $\mathcal{Y}_t$, in comparison with MU and MU+LU. For example, in AVEC'16 (see Fig. 5a), in boxes *2* and *3*, t-LU best captures the whole distribution $\mathcal{Y}_t$, where $\widehat{s}_t$ best resembles s_t. This further highlights the robustness of training on a relaxed s_t through a t-distribution. Backing the quantitative results in Table 1, improvements are more evident in MSPConv, noted from boxes *2* and *3* in Fig. 5c). Crucially, along with the $\widehat{s}_t$ improvements by t-LU, notable improvements are also seen on mean estimates $\widehat{m}_t$.

For *valence*, figures 5b and 5d show that the proposed t-LU evidently improves on mean estimates $\widehat{m}_t$ on both datasets, with only small improvements on standard deviation estimates $\widehat{s}_t$. This can be seen for instance in box *1* of Fig. 5b. Hence, capturing s_t in valence by only relying on audio is a challenging task, and more complex in datasets such as the MSPConv where some frames have a very high subjectivity (see log histograms in Fig.4). It is a common trend in the literature that the audio modality insufficiently explains ground-truth valence m_t [13], [60], and this trend is even more challenging for modeling s_t in valence.

5.3 Analysis on training loss curve

To further study the advantages of the proposed t-distribution $\mathcal{L}_{KL}$ (16) during the training phase, we compare the testing loss curve of (16) with the Gaussian $\mathcal{L}_{KL}$ in MU+LU (12). The comparisons can be seen in Figure 6.

Figure 6 illustrates two crucial advantages of the proposed t-distribution $\mathcal{L}_{KL}$ loss term (16) during training in both datasets. Firstly, we see that in the initial epochs, before epoch 20, the proposed loss converges quicker than the Gaussian $\mathcal{L}_{KL}$ (12). This is the result of the proposed $\mathcal{L}_{KL}$ (16) loss term which penalizes more for lower s_t values, in comparison to the Gaussian $\mathcal{L}_{KL}$ (12) (see Sec.



(a) For arousal, in **AVEC'16** dataset.

(b) For valence, in **AVEC'16** dataset.

(c) For arousal, in **MSPConv** dataset.

(d) For valence, in **MSPConv** dataset.

Fig. 7: Cross-corpora evaluations, for *Match* and *Unmatch* conditions in terms of $\mathcal{L}_{\text{ccc}(m)}$ and $\mathcal{L}_{\text{KL}}$.

3.3.3), thereby achieving faster convergence. Secondly, during the later epochs, after epoch 70, the *Gaussian* $\mathcal{L}_{KL}$ (12) shows signs of overfitting, which is more evident in the MSPConv dataset. However, at the same time, the proposed t-distribution $\mathcal{L}_{KL}$ (16) converges to the best minima during the later epochs. For instance, in MSPConv, the proposed (16) achieves minima $\mathcal{L}_{KL}$ at epoch 86, with $\mathcal{L}_{KL}$ of 0.357 for arousal and 0.373 for valence, while the Gaussian achieves a minima well before the later epochs, at epoch 35, with $\mathcal{L}_{KL}$ of 0.369 for arousal and 0.411 for valence. The proposed $\mathcal{L}_{KL}$ (16) is free from overfitting in the later stages of training and also learns the optima at this stage, noticed across two datasets. This behaviour can be attributed to the nature of the proposed $\mathcal{L}_{KL}$ (16) which promotes the model to learn a more relaxed s_t, thereby introducing more regularization to the model, preventing overfitting and converging on an improved s_t.

5.4 Cross-corpora evaluation

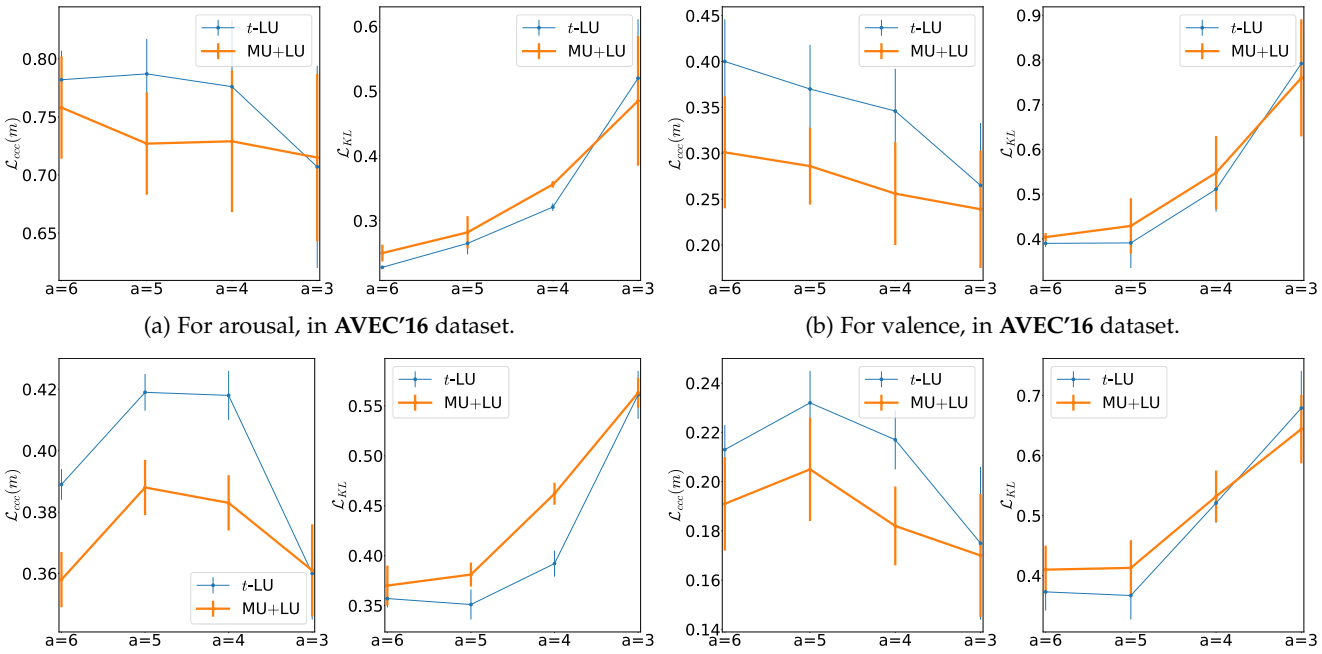
To validate the robustness and generalisation capabilities of the models, we performed cross-corpora evaluations. In Figure 7, results are presented in terms of $\mathcal{L}_{\text{ccc}}(m)$ and $\mathcal{L}_{KL}$, under two conditions. The *Match* condition where the train and the test partitions come from the *same* dataset, and the *Unmatch* condition where the *train* partition is from a *different* dataset. Apart from the dataset size, other dataset-specific factors, such as population demographics and social context, severely challenge the cross-corpora performances because human behaviour varies across group-sizes [36], [61] and social contexts [32]. Crucial differences exist between the AVEC'16 and MSPConv datasets. While the social context of AVEC'16 is a *dyadic* interaction in a *virtual* setting, MSPConv comprises of *larger groups* in a *face-to-face* setting. Moreover, AVEC'16 was collected from *French*-speaking persons, while MSPConv from *English*-speaking persons.

Figure 7 illustrates that the proposed t-LU achieves the best cross-corpora performances on both datasets, and MU with the second best performances. Under the *Unmatch* condition, for *arousal* in AVEC'16 (see Fig. 7a), t-LU achieves 0.421 $\mathcal{L}_{\text{ccc}}(m)$ and 0.409 $\mathcal{L}_{KL}$, while MU achieves 0.342 and 0.490, respectively. Similarly, in MSPConv (see Fig. 7c), t-LU achieves 0.260 $\mathcal{L}_{\text{ccc}}(m)$ and 0.600 $\mathcal{L}_{KL}$, while MU achieves 0.216 and 0.655, respectively.

All models degrade in performance from the *Match* to *Unmatch* conditions. For both arousal and valence, across datasets and metrics, t-LU achieves the *least degrade percentage* while the MTL PU results in the *highest* degrade. For instance, in AVEC'16, in terms of *arousal* mean-estimates $\mathcal{L}_{\text{ccc}}(m)$ (see Fig. 7a), t-LU achieves the least degradation of 46% and MTL PU degrades the most with 61%. Similarly, for *valence* (see Fig. 7b), t-LU degrades least with 53%, and MTL PU degrades the most with 62%. This further emphasises on the robustness of the proposed t-LU model and clearly highlights the lack of robustness of the MTL PU baseline. The MTL PU which achieves the best $\mathcal{L}_{\text{ccc}}(m)$ for valence on the AVEC'16 (see Table 1a), degrades the most on cross-corpora evaluations. This drawback of the MTL PU baseline stems from the dataset-dependent tuning of loss function that it relies on. The proposed t-LU is free from such dataset-dependent tuning and hence more robust. The degrade *percentage* in $\mathcal{L}_{KL}$ is not comparable as the scale of the measure is not linear (depicted in Fig. 2). Also notable is that, for all models, the degrade percentage is larger for valence than for arousal.

5.5 Impact of number of annotations a available

In Sec. 5.1, we noted the benefits of modeling $\mathcal{Y}_t$ as a t-distribution, with *six* available annotations. To capitalise on



(a) For arousal, in **AVEC'16** dataset.

(b) For valence, in **AVEC'16** dataset.

(c) For arousal, in **MSPConv** dataset.

(d) For valence, in **MSPConv** dataset.

Fig. 8: Impact of number of annotations available $a = 6, 5, 4, 3$ on $\mathcal{L}_{\text{ccc}(m)}$ and $\mathcal{L}_{\text{KL}}$.

TABLE 2: Ablation study results of the *t-LU* model, on the AVEC'16 [37] and MSP-Conversation [31] datasets. Modules included in the ablation study are the Uncertainty Layer (BBB), the end-to-end Feature Extractor (E2E), and the Label Distribution Learning loss (KL). ✓ denotes the *inclusion* of the respective module, and × its *omission*. **Bold** results denote the *best two* results for a particular metric, and underline denotes the *least two*. $*$ indicates statistically significant better results over non-bold results. Absence of $*$ indicates that the improvements are not statistically significant.

	Modules			Arousal			Valence		
	E2E	BBB	KL	$\mathcal{L}_{\text{ccc}}(m)\uparrow$	$\mathcal{L}_{\text{ccc}}(s)\uparrow$	$\mathcal{L}_{\text{KL}}\downarrow$	$\mathcal{L}_{\text{ccc}}(m)\uparrow$	$\mathcal{L}_{\text{ccc}}(s)\uparrow$	$\mathcal{L}_{\text{KL}}\downarrow$
AVEC'16	✓	✓	✓	**0.782***	**0.381***	**0.228***	0.400	0.050	**0.390***
	✓	✓	×	0.743	0.356	0.412	0.329	**0.054**	0.594
	✓	×	✓	0.704	0.315	0.299	0.373	0.039	0.426
	✓	×	×	0.721	**0.392***	0.512	**0.401**	**0.064**	0.863
	×	✓	✓	**0.772***	0.330	**0.276***	0.366	0.033	**0.411**
	×	✓	×	0.758	0.329	0.446	0.330	0.050	0.601
	×	×	✓	0.716	0.329	0.318	0.381	0.039	0.446
	×	×	×	0.740	0.310	0.776	**0.420***	0.032	0.960
MSPConv	✓	✓	✓	**0.389***	**0.118***	**0.357***	**0.213***	**0.032**	**0.373***
	✓	✓	×	0.286	0.097	0.412	0.180	0.029	0.495
	✓	×	✓	0.163	0.051	0.515	0.122	0.009	0.537
	✓	×	×	0.271	0.100	0.489	0.174	0.012	0.593
	×	✓	✓	**0.401***	0.056	**0.392***	**0.230***	0.017	**0.391***
	×	✓	×	0.308	0.078	0.551	0.181	0.026	0.416
	×	×	✓	0.247	0.040	0.490	0.140	0.005	0.549
	×	×	×	0.296	**0.107**	0.527	0.181	**0.030**	0.560

the benefits of the t-distribution t-LU over the Gaussian MU+LU, especially when *fewer annotations* are available, we performed experiments by varying a and thereby the degrees of freedom ν. The results are presented in Figure 8, under 4 settings, $a = 3$, $a = 4$, $a = 5$, and, $a = 6$. Annotations were ignored to achieve conditions of $a \leq 5$. The order of annotation to be ignored was handled based on Pearson's correlations measure. For instance, under setting $a = 4$, annotations from two annotators, with the least inter-annotator correlation, for the whole audio file were ignored to model ground-truth annotation distribution $\mathcal{Y}_t$.

Figure 8 shows that, *especially when* $a \geq 4$, the t-distribution based t-LU shows superior performance over the Gaussian MU+LU on both datasets. Crucially, the improvements are larger and more evident when $a = 4$ and $a = 5$ than when $a = 6$. In the case of $\mathcal{L}_{\text{CCC}}(m)$, a non-monotonic behavior with the available number of annotations is notable; $\mathcal{L}_{\text{CCC}}(m)$ initially increases from $a = 6$ to $a = 5$ and subsequently decreases with reducing annotations $a \leq 4$. The initial increase is noticed as annotations are ignored in the order of reducing Pearson's correlation, hence we can expect better consensus in m_t for $a = 5$ than $a = 6$. The subsequent decrease can be associated with the reduced number of annotations to model a stable distribution $\mathcal{Y}_t$.

This emphasises the advantage of t-distribution over the Gaussian with increasing inter-annotator correlation and reducing number of available annotations. In the case of $a = 3$, the performance of t-LU drops below that of the Gaussian MU+LU, as t-LU becomes highly uncertain with only 3 annotations because of the large relaxation on s_t introduced by the scaling in Equation 7 (see Supplementary Sec. 2 for theoretical analysis). This behaviour is similar to the t-test calculation, where models become more uncertain with reducing ν. For modeling emotion annotations as a distribution and uncertainty modeling, we therefore recommend the t-distribution over the Gaussian when more than 3 annotations are available. Noting that both t-distribution and Gaussian drop in performances with only 3 annotations, we suggest collecting at least 4 annotations to obtain a reliable annotation distribution and its ground-truth consensus.

5.6 Ablation study

The proposed end-to-end label uncertainty model has three essential modules, namely 1) feature extractor, 2) uncertainty layer, and, 3) label uncertainty loss. To understand the modules' specific contributions, we perform an ablation study and present its results in Table 2. In case of the feature extractor, *E2E*, ✓ denotes usage of an end-to-end feature extractor and $\times$ the hand-crafted features [20], [62]. In case of the uncertainty layer, *BBB*, ✓ denotes the usage of the BBB-based uncertainty layer and $\times$ the MTL-based s_t estimator. For label uncertainty loss, *KL*, ✓ denotes using $\mathcal{L}_{KL}$ loss and $\times$ denotes usage of $\mathcal{L}_{\text{ccc}}(s)$ loss.

Table 2 firstly shows that end-to-end models achieve better uncertainty estimates than hand-crafted feature models. For instance, in AVEC'16, *E2E* based BBB-KL model achieves 0.381 $\mathcal{L}_{\text{ccc}}(s)$ and 0.228 $\mathcal{L}_{KL}$, improving over hand-crafted features based BBB-KL model which achieves 0.330 and 0.276, respectively. Similarly, in the larger and more complex MSPConv, the E2E based BBB-KL model achieves the best uncertainty estimate performances, against all other models in comparison, with 0.118 $\mathcal{L}_{\text{ccc}}(s)$ and 0.357 $\mathcal{L}_{KL}$. This trend is inline with literature that suggests end-to-end learning, that learns emotion representations in a data-driven manner, for uncertainty modeling [14]. Secondly, the combination of BBB-based uncertainty layer and KL-based loss term (BBB + KL) results in improved performances in both mean and uncertainty estimates, recommending the combination of BBB-layer and KL-loss for label uncertainty modeling in SER. The performance of BBB-layer with a $\mathcal{L}_{\text{ccc}}(s)$ loss term degrades performance across metrics. An intuition behind this is that KL-based *distribution* loss is apt for optimizing the weight *distributions* $P(w|\mathcal{D})$, rather than a loss with only optimizes for s_t of label distribution. Thirdly, across datasets, for both arousal and valence, the KL-based loss term contributes to the improvement of both uncertainty and mean estimates, as the KL loss jointly optimizes for m_t and s_t. For instance, in terms of arousal, the inclusion of KL loss to the E2E+BBB architecture results in a 5% improvement on mean estimates $\mathcal{L}_{\text{ccc}}(m)$ in AVEC'16 and 26% in MSPConv. At the same time, improvements on uncertainty estimates are also noted, 7% improvement of $\mathcal{L}_{\text{ccc}}(s)$ in AVEC'16 and 18% in MSPConv.

Finally, MTL-based s_t estimating model achieves the best $\mathcal{L}_{\text{ccc}}(m)$ performance for valence, but only in AVEC'16 (see last row in Table 2). However, in MPSConv, the proposed BBB+KL based models achieve better results. This improvement, noted only for valence in the AVEC'16, again stems from the dataset-dependent tuning of the loss that is required by MTL-based s_t estimating models (see Sec. 5.1.1). However, this tuning also results in MTL-based s_t estimating models losing their robustness and generalisation capabilities, as shown in cross-corpora evaluations (see Sec. 5.4). Moreover, the MTL-based uncertainty models collapse when not trained on $\mathcal{L}_{\text{ccc}}(s)$ loss, and are not capable of distribution learning using the $\mathcal{L}_{KL}$ loss. Overall, these trends suggest that BBB-based $\mathcal{Y}_t$ learning models are to be preferred over MTL-based s_t estimating models for label uncertainty modeling in SER.

6 Conclusion

We introduced an end-to-end BNN capable of modeling emotion annotations as a label distribution, thereby accounting for the inherent subjectivity-based label uncertainty. In the literature, emotion annotations are commonly modeled using a Gaussian distribution or a histogram representation, however with assumptions based on only limited annotations. In contrast, in this work, we modeled ground-truth emotion annotations as a Student's t-distribution, which also accounts for the number of annotations available. Specifically, we derived a t-distribution based KL divergence loss that, for limited and sparse annotations, produces robust mean estimates and standard deviation estimates that well capture the outliers. We showed that the proposed t-distribution loss term leads to training on a relaxed standard deviation, which is adaptable with respect to the number of annotations available. We validated our approach on two publicly available in-the-wild datasets. Quantitative analysis of the results showed that our proposed approach achieved state-of-the-art results for mean and uncertainty estimations, in terms of both CCC and KL divergence measures, which were also consistent for cross-corpora evaluations. By analysing the loss curves, we showed that the proposed loss term yields faster and improved convergence, and is less prone to overfitting than the Gaussian loss term. Our results further revealed that the advantage of t-distribution over the Gaussian grows with increasing inter-annotator correlation and decreasing numbers of available annotations. Finally, our ablation study suggests that, for modeling label uncertainty in SER, BBB-based label distribution learning models are to be preferred over estimating standard deviation as an auxiliary task.

6.1 Limitations and Future Avenues

In our work, we modeled the emotion annotations as a label distribution using a BNN. However, the BNN introduced here, both MU+LU and t-LU, *jointly* captures the two types of uncertainty— model and label uncertainty. In future work, it would be interesting to focus on disentangling the two types of uncertainty for reliable label uncertainty aware SER systems. One possible way to achieve this concerns *Prior Networks* (PNs) [63], a variant of BNNs, which could be employed to exclusively capture the label uncertainty. PNs do this by parameterizing a prior distribution over predictive label distributions.

This work specifically focused on modeling emotion annotations in a time- and value-continuous manner. In future work, the proposed methodology can be directly extended to model emotion annotations at the utterance-level, as opposed to time-continuous annotations, by simply adding a pooling layer to the feature extractor. However, the method cannot be directly extended to modeling discrete emotion annotations (e.g., classification tasks). Note that the model architecture introduced here (Fig. 1) can be modified to classify discrete emotion labels, but the introduced label uncertainty loss (16) operates only on value-continuous annotations samples. To further extend the introduced loss function for classification tasks, future work may focus on the *discrete* variant of t-distributions. In that case, similar to the loss function derivation in Sec. 3.3.2, KL divergence loss for *discrete* t-distributions would need to be derived.

While the proposed model achieved significantly better state-of-the-art performances in terms of the arousal dimension of emotion across datasets, in one of the datasets (AVEC'16) it did not achieve state-of-the-art performance in terms of *valence*. Note however that state-of-the-art valence performance was achieved in the more complicated MSPConv dataset. It is well documented in the literature that the audio modality insufficiently explains the valence dimension of emotions [55]. This is likely also the reason why the best performing t-LU model, in terms of valence in tables 1a, 1b, and 2, did not achieve statistical significance in some of the metrics despite its improved performance. To overcome this drawback, future work could also include the video and semantic modalities in the feature extractor module, thereby achieving multimodality.

REFERENCES

[1] G. A. Van Kleef, "How emotions regulate social life: The emotions as social information (easi) model," *Current directions in psychological science*, vol. 18, no. 3, pp. 184–188, 2009.

[2] B. W. Schuller, "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.

[3] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[4] D. Dukes, K. Abrams, R. Adolphs, M. E. Ahmed, A. Beatty, K. C. Berridge, S. Broomhall, T. Brosch, J. J. Campos, Z. Clay *et al.*, "The rise of affectivism," *Nature Human Behaviour*, pp. 1–5, 2021.

[5] K. Sridhar, W.-C. Lin, and C. Busso, "Generative approach using soft-labels to learn uncertainty in predicting emotional attributes," in *IEEE Int. Conf. on Affective Comp. and Intelligent Interaction*, Virtual Event, Oct. 2021, pp. 1–8.

[6] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[7] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Jan. 2005, pp. 381–385.

[8] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, pp. 120–136, 2013.

[9] J. Han, Z. Zhang, Z. Ren, and B. Schuller, "Exploring perception uncertainty for emotion recognition in dyadic conversation and music listening," *Cognitive Computation*, vol. 13, Mar. 2021.

[10] K. Sridhar and C. Busso, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Barcelona, Spain, May 2020.

[11] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Calgary, Canada, Apr. 2018.

[12] J. Huang, Y. Li, J. Tao, Z. Lian, and J. Yi, "End-to-end continuous emotion recognition from video using 3D ConvLSTM networks," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Apr. 2018.

[13] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Information Fusion*, vol. 68, pp. 46–53, 2021.

[14] S. Alisamir and F. Ringeval, "On the evolution of speech representations for affective computing: A brief history and critical overview," *IEEE Signal Proc., Magazine*, vol. 38, pp. 12–21, 2021.

[15] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Advances in Neural Inf. Proc. Sys., NeurIPS*, vol. 30, Dec. 2017.

[16] R. Zheng, S. Zhang, L. Liu, Y. Luo, and M. Sun, "Uncertainty in Bayesian deep label distribution learning," *Applied Soft Computing*, vol. 101, Mar. 2021.

[17] M. K. Tellamekala, T. Giesbrecht, and M. Valstar, "Dimensional affect uncertainty modelling for apparent personality recognition," *IEEE Tran. on Affective Computing*, Jul. 2022.

[18] J. Liu, J. Paisley, M.-A. Kioumourtzoglou, and B. Coull, "Accurate uncertainty estimation and decomposition in ensemble learning," in *Advances in Neural Inf. Proc. Sys., NeurIPS*, Vancouver, Dec. 2019.

[19] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, and O. Ronneberger, "A probabilistic U-Net for segmentation of ambiguous images," in *Advances in Neural Inf. Proc. Sys., NeurIPS*, Montreal, Canada, Dec. 2018.

[20] M. K. Tellamekala, E. Sanchez, G. Tzimiropoulos, T. Giesbrecht, and M. Valstar, "Stochastic Process Regression for Cross-Cultural Speech Emotion Recognition," in *Interspeech*, Brno, Sep. 2021.

[21] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. Rezende, and S. A. Eslami, "Conditional neural processes," in *Int. Conf. Machine Learning (ICML), Stockholm, Sweden*, Jul. 2018.

[22] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Int. Conf. Machine Learning (ICML), New York City, NY, USA*, Jun. 2016.

[23] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Int. Conf. Machine Learning (ICML)*, Lille, France, Jul. 2015.

[24] H. Fang, T. Peer, S. Wermter, and T. Gerkmann, "Integrating statistical uncertainty into neural network-based speech enhancement," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Singapore, Jan. 2022.

[25] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.

[26] H.-C. Chou, W.-C. Lin, C.-C. Lee, and C. Busso, "Exploiting annotators' typed description of emotion perception to maximize utilization of ratings for speech emotion recognition," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Singapore, Jan. 2022.

[27] N. M. Foteinopoulou, C. Tzelepis, and I. Patras, "Estimating continuous affect with label uncertainty," in *IEEE Int. Conf. on Affective Comp. and Intelligent Interaction*, Virtual Event, Oct. 2021.

[28] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[29] S. Kotz and S. Nadarajah, *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.

[30] C. Villa and F. J. Rubio, "Objective priors for the number of degrees of freedom of a multivariate t distribution and the t-copula," *Computational Statistics & Data Analysis*, vol. 124, pp. 197–219, 2018.

[31] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Interspeech*, Shanghai, China, Oct. 2020.

[32] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Tran. on Affective Computing*, vol. 10, no. 4, pp. 471–483, Dec. 2019.

[33] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, Shanghai, China, Apr. 2013.

[34] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment re-

search in the wild," *IEEE Trans., on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 1022–1040, 2019.
[35] N. Raj Prabhu, C. Raman, and H. Hung, "Defining and Quantifying Conversation Quality in Spontaneous Interactions," in *Comp. Pub. of 2020 Int. Conf. on Multimodal Interaction*, Sep. 2020.
[36] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, Nov. 2009.
[37] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge," in *Proc., of the 6th Int., Workshop on Audio/Visual Emotion Challenge*, New York, NY, USA, 2016.
[38] N. Raj Prabhu, G. Carbajal, N. Lehmann-Willenbrock, and T. Gerkmann, "End-to-end label uncertainty modeling for speech-based arousal recognition using Bayesian neural networks," in *Interspeech*, Incheon, Korea, September 2022.
[39] N. Raj Prabhu, N. Lehmann-Willenbrock, and T. Gerkmann, "Label uncertainty modeling and prediction for speech emotion recognition using t-distributions," in *IEEE Int. Conf. on Affective Comp. and Intelligent Interaction*, Nara, Japan, Oct. 2022.
[40] M. Abdelwahab and C. Busso, "Active learning for speech emotion recognition using deep neural network," in *IEEE Int. Conf. on Affective Comp. and Intelligent Interaction*, Cambridge, UK, Sep. 2019.
[41] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.
[42] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, ""of all things the measure is man" automatic classification of emotions and inter-labeler consistency," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Philadelphia, USA, 2005.
[43] H. M. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *IEEE Int., Joint Conf., on Neural Networks (IJCNN)*, Vancouver, Canada, Jul. 2016.
[44] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," in *Interspeech*, Graz, Sep. 2019.
[45] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proc., of the 25th ACM Int. Conf. on Multimedia*, Mountain View, USA, Oct. 2017.
[46] T. Dang, V. Sethu, and E. Ambikairajah, "Dynamic multi-rater gaussian mixture regression incorporating temporal dependencies of emotion uncertainty using kalman filters," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Calgary, Canada, Apr. 2018.
[47] G. Rizos and B. Schuller, "Modelling sample informativeness for deep affective computing," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Brighton, UK, May 2019.
[48] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability & statistics for engineers and scientists*. Pearson Education, 2007.
[49] C. Villa and S. G. Walker, "Objective prior for the number of degrees of freedom of at distribution," *Bayesian Analysis*, vol. 9, no. 1, pp. 197–220, 2014.
[50] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, 3rd ed., ser. 5. MIT Press, Jul. 2016, vol. 4, ch. 3, pp. 51–77.
[51] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.
[52] K. P. Murphy, *Machine learning : a probabilistic perspective*. Cambridge, USA: MIT Press, 2012.
[53] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Inf. Proc. Sys., NeurIPS*, Vancouver, Dec. 2019.
[54] D. Rees, "Essential statistics," *American Statistician*, vol. 55, 2001.
[55] P. Tzirakis, A. Nguyen, S. Zafeiriou, and B. W. Schuller, "Speech emotion recognition using semantic information," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Toronto, Jun. 2021.
[56] S. Butterworth *et al.*, "On the theory of filter amplifiers," *Wireless Engineer*, vol. 7, no. 6, pp. 536–541, 1930.
[57] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex fourier series," *Mathematics of computation*, vol. 19, no. 90, pp. 297–301, 1965.
[58] R. Gupta, K. Audhkhasi, Z. Jacokes, A. Rozga, and S. Narayanan, "Modeling multiple time series annotations as noisy distortions of the ground truth: An expectation-maximization approach," *IEEE Tran. on Affective Computing*, vol. 9, no. 1, pp. 76–89, 2016.
[59] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Tran. on Affective Computing*, vol. 6, no. 2, pp. 97–108, 2014.
[60] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," *IEEE Transactions on Affective Computing*, pp. 1–17, Jun. 2022.
[61] C. Raman, N. Raj Prabhu, and H. Hung, "Perceived conversation quality in spontaneous interactions," *IEEE Tran. on Affective Computing*, pp. 1–13, Jan. 2023.
[62] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Tran. on Affective Computing*, vol. 7, no. 2, pp. 190–202, Jul. 2015.
[63] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," *Advances in Neural Inf. Proc. Sys., NeurIPS*, vol. 31, 2018.

Navin Raj Prabhu received a B.Tech degree in Computer Science from SRM University, India, in 2015, and the MS degree in Computer Science from the Intelligent Systems Department at the Delft University of Technology, Delft, The Netherlands, in 2020. Currently, he is a PhD student at the Signal Processing Lab and Organisation Psychology Lab, University of Hamburg, Hamburg, Germany. His research interests include affective computing, social signal processing, deep learning, uncertainty modelling, speech signal processing, and group affect.

Nale Lehmann-Willenbrock studied Psychology at the University of Goettingen and University of California, Irvine. She holds a PhD in Psychology from Technische Universität Braunschweig (2012). After several years working as an assistant professor at Vrije Universiteit Amsterdam and Associate Professor at the University of Amsterdam, she joined Universität Hamburg as a full professor and chair of Industrial/Organizational Psychology in 2018, where she also leads the Center for Better Work. She studies emergent behavioral patterns in organizational teams, social dynamics among leaders and followers, and meetings at the core of organizations. Her research program blends organizational psychology, management, communication, and social signal processing. She serves as associate editor for the Journal of Business and Psychology as well as for Small Group Research.

Timo Gerkmann (S'08–M'10–SM'15) studied Electrical Engineering and Information Sciences at the Universität Bremen and the Ruhr-Universität Bochum in Germany. He received his Dipl.-Ing. degree in 2004 and his Dr.-Ing. degree in 2010 both in Electrical Engineering and Information Sciences from the Ruhr-Universität Bochum, Bochum, Germany. In 2005, he spent six months with Siemens Corporate Research in Princeton, NJ, USA. During 2010 to 2011 Dr. Gerkmann was a postdoctoral researcher at the Sound and Image Processing Lab at the Royal Institute of Technology (KTH), Stockholm, Sweden. From 2011 to 2015 he was a professor for Speech Signal Processing at the Universität Oldenburg, Oldenburg, Germany. During 2015 to 2016 he was a Principal Scientist for Audio & Acoustics at Technicolor Research & Innovation in Hanover, Germany. Since 2016 he is a professor for Signal Processing at the Universität Hamburg, Germany. His main research interests are on statistical signal processing and machine learning for speech and audio applied to communication devices, hearing instruments, audio-visual media, and human-machine interfaces. Timo Gerkmann serves as an elected member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and as an Associate Editor of the IEEE/ACM Transactions on Audio, Speech and Language Processing. He received the VDE ITG award 2022.

End-to-End Label Uncertainty Modeling in Speech Emotion Recognition using Bayesian Neural Networks and Label Distribution Learning

Supplementary Material

Navin Raj Prabhu[⋆†], Nale Lehmann-Willenbrock[⋆], and Timo Gerkmann[†]

[⋆]Industrial and Organizational Psychology, [†]Signal Processing Lab,
Universität Hamburg, Germany.

1 Choice of hyperparameters

Table S1: List of hyperparameters used in the study.

Module	Hyperparameter	AVEC'16[1]	MSPConv[2]	Source
Feature Extractor	# Conv1D	3	3	Adopted from [3].
	Conv1D filters	[64, 128, 256]	[64, 128, 256]	Adopted from [3].
	Conv1D kernel	[8, 6, 6]	[8, 6, 6]	Adopted from [3].
	Conv1D stride	[1, 1, 1]	[1, 1, 1]	Adopted from [3].
	MaxPool kernel	[10, 5, 5]	[10, 5, 5]	Adopted from [3].
	# LSTM	2	2	Adopted from [3].
	LSTM hidden-size	256	256	Adopted from [3].
	Dropout	$p = 0.5$	$p = 0.5$	Adopted from [3].
Uncertainty Layer	# layers	3	3	Adopted from [3].
	Prior $P(w)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	Adopted from [4].
	μ_w of $P(w\|D)$	$\in [-0.1, 0.1]$	$\in [-0.1, 0.1]$	Tuned using grid-search.
	ρ_w of $P(w\|D)$	$\in [-3, -2]$	$\in [-3, -2]$	Tuned using grid-search.
	BBB window-size b	2 s (50 frames)	4 s (100 frames)	w.r.t time-complexity.
	# forward passes n	30	30	w.r.t time-complexity.
	# annotations ν	6	6	As available in [1] and [2].
Training	Samplerate audio	16 kHz	16 kHz	Adopted from [3].
	Samplerate labels	60 Hz	60 Hz	Adopted from [3].
	Optimizer	ADAM	ADAM	Adopted from [3].
	Learning rate	10^{-4}	10^{-4}	Adopted from [3].
	Batch size	5	20	Adopted from [3] and w.r.t time-complexity.
	Epochs	100	100	Tuned manually based on the training loss curves.

The hyperparameters of the *feature extractor* (e.g. kernel sizes, filters) are adopted from [5]. A similar extractor with the same hyperparameters has been used in several multimodal emotion recognition tasks with state-of-the-art performance [6, 5].

As the *prior distribution* $P(w)$, [4] recommend a mixture of two Gaussians, with zero means and standard deviations as $\sigma_1 > \sigma_2$ and $\sigma_2 \ll 1$, thereby obtaining a spike-and-slab prior with heavy tail and concentration around zero mean. But in our case, we do not need mean-centered predictions, as $\mathcal{Y}$ does not follow such a distribution in both datasets (see Sec. 4 in the paper). In this light, we propose to use a simple Gaussian prior with unit standard deviation $\mathcal{N}(0, 1)$. Moreover, a simple $\mathcal{N}(0, 1)$ prior initialization also makes the proposed model scalable across SER datasets.

The μ_w and ρ_w of the *posterior distribution* $P(w|D)$ are initialized uniformly in the range $[-0.1, 0.1]$ and $[-3, -2]$, respectively. The ranges were fine-tuned using grid search for maximized $\mathcal{L}_{\text{KL}}$. For the AVEC'16 dataset, as the test partition is not publicly available, the fine-tuning of $P(w|D)$ is performed using the

train partition. For the MSPConv dataset, the development partition is used. Also, note that the *posterior distribution* $P(w|D)$ and time-shift for post-processing are the only parameters tuned using the partitions.

It is computationally expensive to sample new weights at every time-step (40 ms) and also the level of uncertainties varies rather slowly. In this light, for the AVEC'16 dataset, we set the *BBB window-size* $b = 2$ s (50 frames). As the MSPConv dataset is comparatively larger, a compromise was made for computational simplicity and $b = 4$ s (100 frames) is used. For median filtering, a window-size of 2 s is used. In this work, we assume a Gaussian on $\widehat{\mathcal{Y}}_t$, and noted previously that $n \geq 30$ is required for the assumption to hold. In this light, and keeping the time-complexity in mind, we fixed $n = 30$.

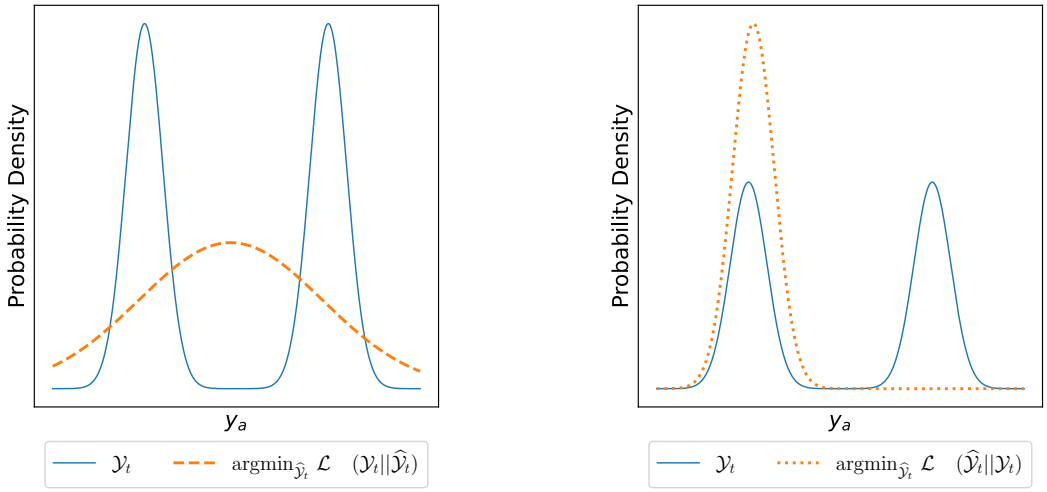
For training, we use the Adam optimizer with a learning rate 10^{-4}. The batch size used was 5 and 20, for AVEC'16 and MSPConv, respectively, with a sequence length of 300 frames, 40 ms each. All models were trained for a fixed 100 epochs. The complete list of hyperparameters used by this work is listed in Table S1.

2 Post-processing

For all the baselines and models proposed in this work, two post-processing techniques are applied, namely, median filtering [3] with window-size same as the BBB window-size b, and time-shifting [7] (with shifts between 0.04s and 10s). To find the best time-shift, a grid-search was performed between 0.04s and 10s using the training partition in AVEC'16 and the development partition in MSPConv. Specifically, the grid-search is performed to maximize $\mathcal{L}_{\text{ccc}}(m)$ metric in the respective partition, and subsequently the best time-shift is used to also recalculate the $\mathcal{L}_{\text{ccc}}(s)$ and $\mathcal{L}_{\text{KL}}$ measures.

The following trends were noticed during the post-processing of accounting for the annotator lag. On one hand, in the *MSPConv* dataset, correction for annotator lag was not required for most of the models and baselines. This is because, as detailed in Sec. 4.1.2, we performed median and low-pass filtering on the continuous annotations, for a uniform sampling rate and to remove periodic distortions noticed in the dataset. These filtering techniques which inherently use sliding-windows might have already filtered out the annotator lags. On average across the baselines and models, correction for a lag of 0.24 was sufficient to achieve the best results. On the other hand, in the *AVEC'16* dataset, on average across the baselines and models, a rather large correction of 1.36s was required to achieve the best results.

3 Mean- and Mode- seeking KL divergence



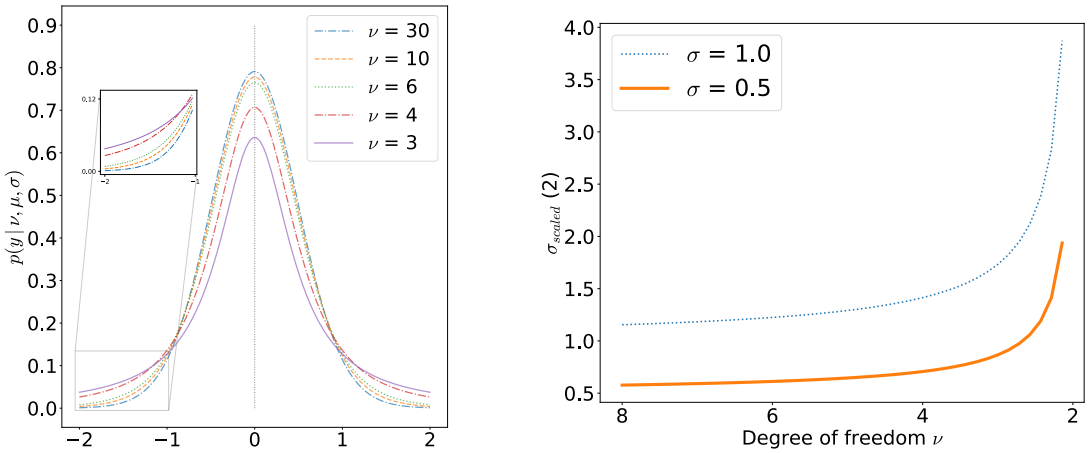
(a) *Mean*-seeking approximation: $\mathcal{L}_{\text{KL}}(\mathcal{Y}_t||\widehat{\mathcal{Y}}_t)$ **(b)** *Mode*-seeking approximation: $\mathcal{L}_{\text{KL}}(\widehat{\mathcal{Y}}_t||\mathcal{Y}_t)$

Figure S1: Comparison between the mean- and mode- seeking approximations of KL divergence $\mathcal{L}_{\text{KL}}$.

The KL divergence $\mathcal{L}_{\text{KL}}$ is asymmetric. We have the choice of minimizing either $\mathcal{L}_{\text{KL}}(\mathcal{Y}_t||\widehat{\mathcal{Y}}_t)$ or $\mathcal{L}_{\text{KL}}(\widehat{\mathcal{Y}}_t||\mathcal{Y}_t)$. In Figure S1, we illustrate the difference between the two choices of approximations: the *mean*-seeking approximation, where the ground-truth distribution $\mathcal{Y}_t$ is followed by its estimate distribution $\widehat{\mathcal{Y}}_t$, and the mode-seeking approximation, where the order is reversed and the estimate $\widehat{\mathcal{Y}}_t$ is followed by the ground-truth $\mathcal{Y}_t$. In case of the *mean*-seeking approximation (Figure S1a), when $\mathcal{Y}_t$ has multiple modes, the estimate $\widehat{\mathcal{Y}}_t$ blurs the modes together by estimating high probability mass on all of them, thereby capturing the whole

distribution [8]. But in case of the *mode*-seeking approximation (Figure S1b), when $\mathcal{Y}_t$ has multiple modes, $\mathcal{L}_{\text{KL}}$ is minimized by fitting on a *single mode*, thereby not capturing the whole distribution [8]. However we argue that, in our case of modeling emotion annotations y_a as a distribution $\mathcal{Y}_t$, we require the estimate distribution $\widehat{\mathcal{Y}}_t$ to capture the whole distribution without fitting on a single mode. Intuitively, when $\widehat{\mathcal{Y}}_t$ is fit on a single mode it fails to produce reliable mean and standard-deviation estimates, a crucial goal in uncertainty modeling for emotion recognition research. Moreover, our preliminary experiments comparing the mean- and mode-seeking approximations also indicated that the mean-seeking approximation tends to achieve better distribution modeling results than the mode-seeking one.

4 Modeling distributions with only 3 samples: Theoretical analysis



(a) Highly-uncertain distribution as ν decreases.

(b) Increasing standard deviation σ_{scaled} as ν decreases

Figure S2: Effect of degree of freedom ν on the scaling of σ and highly-uncertain distribution.

Unlike the Gaussian distribution, the t-distribution also accounts for the number of samples used to model the distribution through the degree of freedom ν included in its probability density function,

$$p(y \mid \nu, \mu, \sigma) = \frac{1}{\mathrm{B}(\frac{1}{2}, \frac{\nu}{2})} \frac{1}{\sqrt{\nu\sigma^2}} \left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}, \tag{1}$$

where B(., .) is the Beta function, for Gamma function Γ, formulated as $B(i,j) = \dfrac{\Gamma(i)\,\Gamma(j)}{\Gamma(i+j)}$. Furthermore, the standard deviation σ of the t-distribution, in (1), takes the scaled form, where σ is scaled using the normality parameter ν:

$$\sigma \sqrt{\frac{\nu}{\nu-2}} \text{ for } \nu > 2. \tag{2}$$

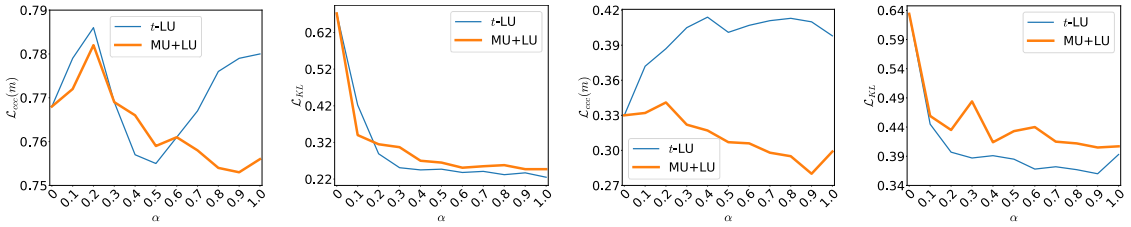
Let us denote this scaled form of the standard deviation as σ_{scaled}. Through this scaling (2), the standard deviation of the t-distribution σ_{scaled} *increases* with decreasing number of annotation samples available to model the distribution [9]. Bishop [10] associates this scaled standard deviation σ_{scaled} towards the increased robustness of the t-distribution towards outliers and sparse distributions. This is also noticed from the results of the experiments presented where the t-distribution is superior in modeling the annotation distribution over the Gaussian. However, a caveat of this σ_{scaled} is that when the number of annotations samples is *less than 4*, the t-distribution associates this as a highly-uncertain distribution with a highly scaled σ. Figure S2 further illustrates the impact of σ_{scaled} on the probability density function of the t-distribution (1). Figure S2a depicts the increasing uncertainty in the t-distribution as the degrees of freedom ν decreases. The zoomed region further highlights the case of $\nu = 3$ (only 3 annotations available) where a relatively high likelihood is associated along the tails, thereby the distribution becomes highly-uncertain. Figure S2b illustrates the increasing scaled standard deviation σ_{scaled} with reducing ν. It is noted here that, for $\nu \leq 3$, the rate of increase in standard deviation further enlarges, thereby explaining the reason why label distribution modeling fails when only 3 annotation samples are available. Note that this highly-uncertain distribution and scaled standard deviation also affects the Kullback–Leibler divergence loss thereby affecting the training process.

5 Effect of α: regularization with label uncertainty loss term $\mathcal{L}_{\mathrm{KL}}$

The proposed end-to-end uncertainty loss is,

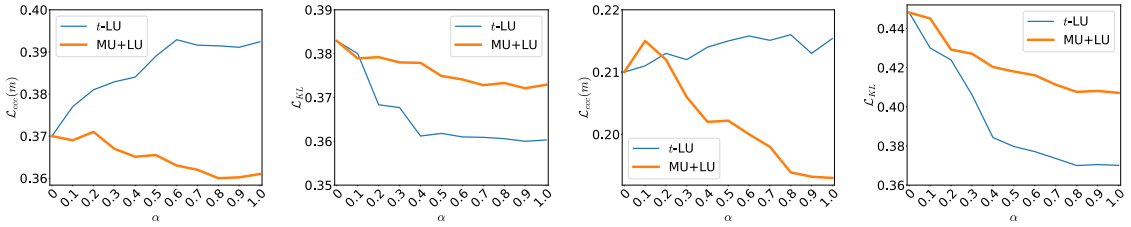
$$\mathcal{L} = (1 - \mathcal{L}_{\mathrm{CCC}}(m)) + \mathcal{L}_{\mathrm{BBB}} + \alpha \mathcal{L}_{\mathrm{KL}}. \tag{3}$$

Intuitively, $\mathcal{L}_{\mathrm{CCC}}(m)$ optimizes for mean predictions m, $\mathcal{L}_{\mathrm{BBB}}$ optimizes for BBB weight distributions, and $\mathcal{L}_{\mathrm{KL}}$ optimizes for the label distribution $\mathcal{Y}_t$. The variable α controls the degree to which the model is regularized on the label uncertainty loss term $\mathcal{L}_{\mathrm{Kl}}$. For $\alpha = 0$, the model only captures model uncertainty (MU). For $\alpha = 1$, the model also captures *label uncertainty* (MU+LU or t-LU). To further understand the effect of the regularization weighting factor α, we performed experiments with varying α from 0 to 1 and with a hop of 0.1. The results of the experiments, in-terms of the $\mathcal{L}_{\mathrm{CCC}}(m)$) and $\mathcal{L}_{\mathrm{KL}}$ metrics, for the AVEC'16 [1] and MSPConv [2] datasets can be seen in Figures S3 and S4, respectively.



(a) $\mathcal{L}_{\mathrm{ccc}}(m)$ in *arousal* **(b)** $\mathcal{L}_{\mathrm{KL}}$ in *arousal* **(c)** $\mathcal{L}_{\mathrm{ccc}}(m)$ in *valence* **(d)** $\mathcal{L}_{\mathrm{KL}}$ in *valence*

Figure S3: Effect of α: regularization with label uncertainty loss term $\mathcal{L}_{\mathrm{KL}}$, in the **AVEC'16** dataset.



(a) $\mathcal{L}_{\mathrm{ccc}}(m)$ in *arousal* **(b)** $\mathcal{L}_{\mathrm{KL}}$ in *arousal* **(c)** $\mathcal{L}_{\mathrm{ccc}}(m)$ in *valence* **(d)** $\mathcal{L}_{\mathrm{KL}}$ in *valence*

Figure S4: Effect of α: regularization with label uncertainty loss term $\mathcal{L}_{\mathrm{KL}}$, in the **MSPConv** dataset.

From figures S3 and S4, we observe the following trends with respect to the regularization factor α. Firstly, with increasing α, as expected, the $\mathcal{L}_{\mathrm{KL}}$ also decreases, with a sharp decrease until $\alpha = 0.3$ and gradually decreasing for $\alpha \geq 0.3$ until it starts plateauing from $\alpha = 0.7$ (seen from figures S3b, S3d, S4b, S4d). This indicates that both the t-LU and MU+LU models reach their maximum capacity in-terms of modeling the label distribution $\mathcal{Y}_t$ with an α greater than 0.7. Furthermore, in-terms of the mean-estimates $\mathcal{L}_{\mathrm{ccc}}(m)$, crucially we note that, with increasing regularization on the label uncertainty loss $\mathcal{L}_{\mathrm{KL}}$, while the t-LU model performance increases with increasing α, the MU+LU model performance drops gradually with increasing α (seen from figures S3a, S3c, S4a, S4c). This behaviour further emphasises that t-LU is free from the compromises MU+LU make on mean-estimates $\mathcal{L}_{\mathrm{ccc}}(m)$ while modeling label uncertainty (detailed in Sec. 5.1.1 in the paper). Similarly to the plateauing of $\mathcal{L}_{\mathrm{KL}}$ for α greater than 0.7, the $\mathcal{L}_{\mathrm{ccc}}(m)$ also starts plateauing from 0.8. Overall, from figures S3 and S4, with respect to both mean-estimate modeling $\mathcal{L}_{\mathrm{ccc}}(m)$ and label distribution modeling $\mathcal{L}_{\mathrm{KL}}$, we recommend using a regularization factor of $\alpha \in [0.8, 1.0]$.

References

[1] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge," in *Proc., of the 6th Int., Workshop on Audio/Visual Emotion Challenge*, New York, NY, USA, 2016.

[2] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Interspeech*, Shanghai, China, Oct. 2020.

[3] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Calgary, Canada, Apr. 2018.

[4] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Int. Conf. Machine Learning (ICML)*, Lille, France, Jul. 2015.

[5] P. Tzirakis, A. Nguyen, S. Zafeiriou, and B. W. Schuller, "Speech emotion recognition using semantic information," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Toronto, Jun. 2021.

[6] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Information Fusion*, vol. 68, pp. 46–53, 2021.

[7] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Tran. on Affective Computing*, vol. 6, no. 2, pp. 97–108, 2014.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, 3rd ed., ser. 5. MIT Press, Jul. 2016, vol. 4, ch. 3, pp. 51–77.

[9] C. Villa and F. J. Rubio, "Objective priors for the number of degrees of freedom of a multivariate t distribution and the t-copula," *Computational Statistics & Data Analysis*, vol. 124, pp. 197–219, 2018.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

3.2 Group-level Affect: Annotations and Multimodal Modeling [P4]

Abstract

Collaborating in a purposive group, whether face-to-face or virtually, involves continuously expressing emotions and interpreting those of other group members. As such, understanding group affect is essential to comprehending how groups interact and succeed in collaborative efforts. In this study, we move beyond individual-level affect and investigate group-level affect—a collective phenomenon that reflects the shared mood or emotions among group members at a particular moment. As the first in the literature, we gather annotations for group-level affective expressions in purposive group interactions using a fine-grained temporal approach (15 second windows) that also captures the inherent dynamics of this collective construct. To this end, we extensively train annotators and develop an annotation procedure specifically tuned to capture the entire scope of the group interaction from one interaction moment to the next. In addition, we model the ebb and flow of group affect by accounting for the underlying convergence (driven by emotional contagion) and divergence (resulting from emotional reactivity) of affective expressions among group members. To capture these interpersonal dynamics, we employ two approaches: (i) extracting synchrony-based handcrafted features from both audio and visual modalities, and (ii) introducing a novel, data-driven graph neural network to model interpersonal dynamics among group members. Our results highlight the advantages of the graph network over the handcrafted features in modeling group affect, while also emphasizing the importance of temporal modeling and incorporating multimodal cues. Additionally, our analysis of affective convergence and divergence reveals that groups tend to diverge in their social signals during neutral collective affect, while exhibiting convergence during more emotionally intense moments. These insights are drawn from comparative results across both modeling techniques.

Reference

N. Raj Prabhu, M. Tsfasman, C. Oertel, T. Gerkmann, and N. Lehmann-Willenbrock, "Dynamics of Collective Group Affect: Group-level Annotations and the Multimodal Modeling of Convergence and Divergence", *Accepted for IEEE Transactions on Affective Computing*, Dec. 2025.

Copyright Notice

Authors' Contributions

This work was a collaboration between the University of Hamburg and the Technical University of Delft.

From the University of Hamburg: Navin Raj Prabhu, as the lead author, led the study, including the initial conceptualization, planning and organizing the annotation collection, algorithm development, neural network training, experimental validation, and manuscript

preparation. Timo Gerkmann offered valuable methodological feedback through discussions and participated in the manuscript review. Nale Lehmann-Willenbrock contributed through early-stage brainstorming to develop the annotation strategy, supported the lead author in organizing and training the annotators, and assisted in reviewing the manuscript and refining the argumentation and overall framing.

From the Technical University of Delft: M. Tsfasman and C. Oertel were responsible for the original collection of the group interaction dataset, which they shared for the purposes of the annotation collection conducted in this study. They also participated in the manuscript review.

Dynamics of Collective Group Affect: Group-level Annotations and the Multimodal Modeling of Convergence and Divergence

Navin Raj Prabhu, Maria Tsfasman, Catharine Oertel, Timo Gerkmann, and Nale Lehmann-Willenbrock

Abstract—Collaborating in a purposive group, whether face-to-face or virtually, involves continuously expressing emotions and interpreting those of other group members. As such, understanding group affect is essential to comprehending how groups interact and succeed in collaborative efforts. In this study, we move beyond individual-level affect and investigate group-level affect—a collective phenomenon that reflects the shared mood or emotions among group members at a particular moment. As the first in the literature, we gather annotations for group-level affective expressions in purposive group interactions using a fine-grained temporal approach (15 second windows) that also captures the inherent dynamics of this collective construct. To this end, we extensively train annotators and develop an annotation procedure specifically tuned to capture the entire scope of the group interaction from one interaction moment to the next. In addition, we model the ebb and flow of group affect by accounting for the underlying convergence (driven by emotional contagion) and divergence (resulting from emotional reactivity) of affective expressions among group members. To capture these interpersonal dynamics, we employ two approaches: (i) extracting synchrony-based handcrafted features from both audio and visual modalities, and (ii) introducing a novel, data-driven graph neural network to model interpersonal dynamics among group members. Our results highlight the advantages of the graph network over the handcrafted features in modeling group affect, while also emphasizing the importance of temporal modeling and incorporating multimodal cues. Additionally, our analysis of affective convergence and divergence reveals that groups tend to diverge in their social signals during neutral collective affect, while exhibiting convergence during more emotionally intense moments. These insights are drawn from comparative results across both modeling techniques.

Index Terms—Group affect, affect dynamics, annotations, convergence, divergence, multimodal analysis, automatic affect recognition

1 Introduction

Group affect is a *collective* social construct that represents the jointly experienced shared mood or emotions that group members hold in common at a given point in time [1], [2]. There is an important conceptual and empirical difference between the affective experiences of individual group members and the collective mood of the group as a whole, not only regarding how they are measured but also in terms of their influence on group outcomes [3]. Following [2], we focus our research on affect at a collective level in *purposive groups*[1]. By purposive group, we mean "an intact social system, complete with boundaries, interdependence for some shared purpose, and differentiated member roles" [4]. This makes purposive groups distinct from spontaneous social gatherings or large groups. Purposive groups are prevalent in organizational settings, tasked with completing a wide variety of assignments across different time frames [2]. Examples of purposive groups include a group of software developers or a group of health workers assembled to brainstorm solutions to a problem. The relevance of collective group affect towards the functioning and outcomes of purposive groups is documented well in the literature [5]–[12]. For example, a widely cited review of the literature discusses the importance of group affect for shaping (i) group member attitudes, (ii) cooperation and conflict resolution, (iii) creativity and decision making, and (iv) group performance [2]. In terms of the factors that give rise to shared group affect, recent qualitative work points to social learning and positive emotional sharing [13].

However, there is a notable dearth of *quantitative* empirical research on *collective group affect* as it emerges and fluctuates during group interactions. Whereas research has extensively examined affect as an individual-level construct [14]–[18], we know much less about group affect as a collective phenomenon (for an overview and detailed critique, see [8]). This is because gathering suitable group interaction data, annotating, and analyzing *collective group affect* and the related social signals is considerably more complex compared to that of individual members' affect [9], [11]. In this work, we address this knowledge gap through two key contributions: (1) *collecting group affect annotations* using an annotation strategy that aligns methodologically with theoretical frameworks from organisational psychology on group affect, and (2) leveraging these annotations to perform *multimodal modeling of dynamic group affect*, capturing the underlying phenomena of convergence and divergence.

- *Navin Raj Prabhu and Timo Gerkmann are with the Signal Processing Lab, University of Hamburg, Germany, 20146. E-mail: navin.raj.prabhu@uni-hamburg.de, timo.gerkmann@uni-hamburg.de*
- *Maria Tsfasman and Catharine Oertel are with the Department of Intelligent Systems, TU Delft, 2628 Delft, The Netherlands. E-mail: m.tsfasman@tudelft.nl, c.r.m.m.oertel@tudelft.nl*
- *Nale Lehmann-Willenbrock is with the Department of Industrial and Organizational Psychology, University of Hamburg, Germany, 20146. E-mail: nale.lehmann-willenbrock@uni-hamburg.de*

This work was funded under the Excellence Strategy of the Federal Government and the Länder, and the project "Mechanisms of Change in Dynamic Social Interaction" (LFF-FV79, Landesforschungsförderung Hamburg).

1. Throughout this work, we use the term "groups" to specifically represent purposive groups.

Collecting annotations of group affect: We identified two key challenges in annotating group affect from a review on prior work in organizational psychology and computer science: (i) capturing fine-grained temporal dynamics [8], [10], [19], [20], and (ii) accounting for social dynamics in complex purposive groups [2], [21].

The limited literature on group affect to date has predominantly relied on annotations of static images [9], [19], [22] or temporally independent video segments [11], [23] *without accounting for any temporal context* [8], [20]. Annotating video segments in a temporally independent manner without accounting for the temporal context is resource efficient, where annotators need not continuously track the group's affective expressions and its associated changes between consecutive segments. However, such an annotation approach is agnostic to the inherently dynamic nature of group affect [2], [10], which limits the resulting empirical contributions and represents a misalignment between the theoretical construct of group affect as a dynamic process and its measurement [7], [24].

Based on our review across disciplines, we observe considerable *theory-method misalignment* between theoretical models of group affect as presented in the organizational behavior literature (e.g., [1], [2]) and the group affect recognition methodologies developed in computer science studies (e.g., [9], [19], [25]). While the theory conceptualizes group affect in complex *purposive group settings* involving interpersonal relationships and dynamics [2], group affect recognition research has focused on much simpler groups that lack social intactness and interdependence towards a shared goal. As one example, consider the groups captured in the Video Group Affect dataset, which includes concerts, silent protests, and fights [11]. Notably, annotating and modeling collective affect in purposive groups is significantly more complex, as it requires capturing the dynamics of interpersonal relationships, unlike simpler non-purposive groups where such relationships are largely absent.

The two challenges above informed the design of our annotation strategy. To address the challenge of capturing *temporal dynamics* in group affect, we conducted pilot studies to iteratively tune the annotation window size, enabling more apt representation of affective fluctuations. To handle the complexity of annotating affect in *purposive groups*, we extensively trained organizational psychology students, ensuring context-aware, high-quality labels. This approach is consistent with prior research emphasizing the need for domain-specific training and the involvement of social psychologists in annotating group-level phenomena [20], [26]. Informed by [27], our training incorporated video markers identified during the pilot study. Together, these efforts ensure that our annotation process is well-aligned with organizational psychology theory, which frames group affect as a dynamic social construct within purposive groups.

Multimodal modeling of dynamic, group-level affect: Annotating dynamic group affect allows for (i) the *development of multimodal predictive models* capable of automatic recognition of group affect with more real-world application possibilities, coping with changing affect over time, and higher robustness [8], and (ii) the *quantitative analysis of convergence and divergence* in affective expressions that shape the dynamics of group affect [28].

Social signals, such as facial expressions [29] and vocal pitch [14], encode micro-level behavioral cues like synchrony and convergence [30], which are fundamental to the development of dynamic interpersonal relationships [30] and higher-order group constructs [26], [31]. In this work, we introduce two complementary methods for capturing these micro-level patterns using multimodal, individual-level social signals. The first approach leverages handcrafted features designed to quantify synchrony and convergence, thereby modeling interpersonal dynamics (e.g., [26], [31]). The second approach employs a novel graph neural network (GNN), inspired by social network theory [32], which represents individuals as nodes and their relationships as edges, allowing it to learn the structure and temporal evolution of social interactions. Unlike handcrafted features, the GNN offers a data-driven framework for modeling dynamic interpersonal relationships.

Barsade [2], in a review of the organizational psychology literature, highlighted the need for fine-grained analyses of the temporal dynamics underlying collective group affect. In particular, the study of *convergence* and *divergence* in affective expressions are ways to study this ebb and flow of group affect. Hareli and Rafaeli [28] proposed a cyclical model where individual affect converges via contagion or diverges through reactivity, creating a feedback loop that shapes group functioning. Recent conceptual work further emphasizes the need to account for social dynamics in that lead to shared affect in groups [3]. Despite its theoretical significance, empirical research on affective convergence and divergence in group interactions remains limited. To address this gap, we conduct a quantitative analysis of affective convergence and divergence using both the handcrafted features and the data-driven representations from the proposed graph network.

To the best of our knowledge, this is the first study to develop predictive models of *dynamic* group affect in complex *purposive* group interactions, and the first to present an empirical analysis of affective *convergence and divergence* in such dynamic settings. For a detailed comparison of our study contribution beyond existing approaches in the literature, see Appendix Table S2.

2 Background and Related Work

2.1 Affect in purposive groups

While individual affect has been extensively studied in the past few decades [14], collective group affect has received considerably less research attention [2], [8], [20]. Empirical research [33] has demonstrated that group-level affect is distinct from individual-level affect and that they are shared among interlocutors in purposive interactions. This has consequences for research design, requiring scholars to pay attention to the actual social interaction dynamics that give rise to group affect (for an overview, see [7]). Understanding group affect is a key element of understanding how groups interact and achieve collaborative performance [2], [7]. This study specifically focuses on the under-researched construct of collective group affect.

To analyze affect in groups, quantification of affect is the first step. Two types of quantification techniques have been widely used in the literature: (1) Ekman's six basic emotions

[34] (e.g., happy, angry) or (2) Russell's circumplex model [35], where affect is quantified using two continuous, bipolar, and orthogonal dimensions of arousal and valence. In recent years, research in social science as well as state-of-the-art datasets for affect recognition in computer science have moved from categorical representations to the circumplex model, owing to the fact that the circumplex model is more suitable for capturing the ambiguous and fuzzy nature of affective expressions (e.g., [10], [15], [36]). In line with this development across disciplines, in this work we rely on the circumplex model to quantify collective group affect.

2.2 Dynamics inherent in group affect

Theorizing on group affect as a dynamic, temporally evolving social construct characterized by bottom-up and top-down processes, while intuitively appealing, has received only limited empirical validation efforts [2]. This is likely due to the challenge of annotating collective group behavior while also accounting for its inherent temporal dynamics [19], [20]. Two key challenges in this space are: (i) the time and cost inefficiencies stemming from the complexity of group constructs [10], [11], [19], and (ii) the need for domain expertise, particularly the involvement of social psychologists, and intensive annotator training [20], [26].

Moreover, the appropriate window size required to capture the dynamics of group-level social constructs remains an open research question across disciplines. Choices regarding the temporal granularity of annotations differ substantially in the prior literature. For example, Mo et al. [37] used 20-second windows, whereas Barsade [38] or Lei et al. [10] used 2-minute windows to capture dynamic changes in group affect. Of note, recent research on *individual*-level affect has moved from segment-level annotations (≈ 5 secs) [17] to much smaller time windows, as small as 10 ms [16], [18]. Annotating *larger* time windows is simpler and more resource-efficient, as annotators need not track micro-level behaviors and events, and require fewer labels for an entire interaction. However, such windows are less suited for capturing the dynamics of group affect, which often shifts and changes from one conversational moment to the next. In contrast, *smaller* time windows better reflect these dynamics but make the annotation process significantly more complex and time-consuming [18].

In this work, we address these challenges by training annotators with a background in social psychology to attentively track the full scope of group interactions and annotate for dynamic group affect. This training enables annotators to incorporate both top-down and bottom-up processes that underlie the emergence of group-level affect. To aptly capture the dynamic fluctuations of group affect, we eploy a window size that is iteratively tuned with respect to the construct and the social setting at hand.

2.3 Automatic recognition of dynamic group affect

Dhall et al. [9] were among the first to explore automatic group affect recognition using static web images of non-purposive groups (e.g., concerts, silent protests [39]), annotated with six levels of happiness intensity. This foundational work was extended by [25] and [40], and later evolved into a multimodal framework through the efforts of Sharma et al. [11], [41]. To promote further progress in the field, a series of benchmark challenges were introduced [22], [23], [39], and [42] open-sourced an interactive, multimodal group affect analysis toolkit for use in diverse social settings.

More recently, GNNs have gained attention for this purpose due to their ability to model relational structures [20]. Early GNN-based approaches [19], [43] primarily focused on static group images, capturing only spatial relationships among individuals and their environments, without the temporal context. Addressing this gap, Wang et al. [44] introduced a unimodal, video-based GNN that represents individuals as nodes and includes pseudo-nodes to aggregate spatial and temporal features between neighboring time frames, within a fixed-length, temporally independent video segment. Of note, the model is limited to fixed-duration inputs and does not scale well to longer temporal contexts, as the GNN's adjacency matrix expands proportionally with input duration and the growing number of pseudo-nodes.

Despite recent progress, a key limitation remains in group affect recognition research: insufficient temporal modeling—most existing methods treat video segments as temporally independent, overlooking how group affect unfolds over time. To address this, propose a GNN-based model that represents individuals as nodes initialized with multimodal, individual-level social cues, and with edges trained to capture interpersonal relationships. Subsequently, a long short-term memory (LSTM) layer operates on the edge representations to model temporal dynamics between consecutive video segments. Leveraging the collected, temporally evolving annotations, the model captures fluctuations in group affect across variable-length interactions—integrating both bottom-up and top-down processes, thus aligning computational modeling more closely with established psychological theory [2], [8], [10].

2.4 Fine-grained analyses of dynamic group affect

The investigation of affective convergence and divergence allows insights into collective group-level affect and its underlying dynamics [28]. Organizational psychology literature [21], [28] identifies emotional contagion and reactivity as key mechanisms through which convergence and divergence emerge within social groups. These processes involve one person or group influencing the emotions or behaviors of others through the conscious or unconscious induction of emotional states and behavioral responses [28], [45]. For emotional contagion to occur, individuals must first express their emotional states through observable social behaviors, which others may then mirror (i.e., convergence) or contrast with (i.e., divergence). Capturing this dynamic, reciprocal exchange of micro-level behaviors requires the extraction of *synchrony* and *convergence* features, as outlined in [30]. For example, group members may express more activated speech adapting to other group members' level of activation (emotional contagion), or they might gasp in response to a group member's ill-fated story (emotion reactivity).

To capture these affective contagion and reactivity within groups, we employ graph *attention* networks (GATs), a variant of GNNs that learn attention weights between interlocutors represented as nodes. Alongside an analysis based on handcrafted synchrony and convergence features, we

examine the learned GAT attention weights to gain insight into convergence and divergence processes.

3 Dynamic Group Affect: Conceptualization and Annotation

The experimental workflow of this research comprises two pipelines: (1) the Annotation Strategy, used to collect group affect annotations, and (2) the Modeling process, which leverages these annotations to analyze dynamic group affect. An illustration of this workflow is in Appendix Figure S2. This section focuses on the first pipeline, detailing the annotation strategy employed.

Barsade and Knight [2] defined group affect as the amalgamation of group members' affective states and the mutual influence of a group's affective context. Recently, theorists have expanded on the temporal dynamics of group affect, showing how momentary individual- and group- affective experiences become inputs for future group affective experiences [28], [46]. In line with these organization psychology theories on group affect [2], [7], [28], [46], we conceptualize group affect as a dynamic and continually evolving social phenomenon in groups. Specifically, we define *group affect* as the collective affective state of the group, which is the amalgamation of group members' affective states expressed during group interactions (i.e., a *bottom-up* process) and which in turn affects future affective experiences of the group (i.e., *top-down* influence). Of note, this amalgamation requires individual behavior to be expressed in terms of social signals which can be annotated by external annotators.

3.1 Dataset: MEMO

To investigate dynamic group affect in the context of a purposive group interactions, we drew suitable data from the multimodal longitudinal meeting corpus (MeMo) [47], consisting of 45 unscripted video-call discussions in groups of three to six participants. The MeMo corpus, with its setup of purpose-driven group discussions and longitudinal recordings, is an ideal dataset for studying group affect in purposive groups. To achieve maximum affect elicitation, we selected those MeMo group discussions that focused on group members' experiences during the COVID-19 pandemic, recorded in the year 2021. A group discussion on COVID-19 is likely to evoke strong emotional responses due to several factors, including polarized opinions, economic impact, hope and resilience, loss and grief, and shared experiences. The recorded group interactions lasted for approximately 45 mins (average duration of 41 mins and 35 secs; with a standard-deviation of 7 mins and 30 secs). As a longitudinal meeting corpus, participants were divided into 15 groups, and each group met 3 times over the course of 2 weeks using a video-conferencing software. At the start, the participants were reported to be complete strangers, having never met each other before the first session.

The interactions were guided by professional facilitators in order to promote active discussions among the group members. The selected facilitators were experienced in moderating meetings and facilitating creative sessions. To maximize the diversity of in-group opinions, participants were recruited from various COVID-19 affected demographics in every group, e.g., parents with young children, adults of age 50+, students, (ex)-business owners. Overall, 15 groups, totaling 53 participants (25 Male, 28 Female; 18-76 years old) and 4 moderators (3 Male, 1 Female; 24-45 years old) took part in the interactions collected as part of MeMo.

To facilitate future research on these novel group affect annotations, we commit to making the annotations public along with the MeMo corpus [47]. Initially, the annotations will be publicly accessible following the review process and can be obtained by agreeing to the end-user agreement[2]. It is important to note that the annotations can be released immediately, as they have already been anonymized using pseudonyms for both annotators and interlocutors and do not include any sensitive personal information. Conversely, the raw behavioral group interaction data recorded in the MeMo corpus contains sensitive and private information shared by the interlocutors [47]. To protect the data of the interlocutors, a thorough review process is currently being conducted to eliminate any sensitive or potentially private information. Following this process, the MeMo corpus will also be made publicly available in the same data repository as the annotations (please see [47] for more details).

For this work, we used only the spontaneous interaction segments within the MeMo corpus. All interactions had eye-gaze calibration and administrative tasks at the beginning and at the end, respectively. Furthermore, in some interactions, participants were late to join the interaction. We cropped these segments out of all the interaction videos for our annotation procedure. The timestamps of these segments were manually annotated by the lead author.

3.2 Annotators

In [26], [48], efforts to annotate conversation quality at both individual and group levels reveal that group-level constructs are more complex to annotate and typically have lower inter-annotator agreement than individual-level constructs. The importance of involving social psychologists in the annotation of group affect is further underscored in [20].

Following this, instead of relying on naive annotators, we trained annotators for the task of annotating dynamic group affect. This approach helped us overcome several limitations regarding the annotation of group affect in the prior literature. These include: (i) accurately capturing the dynamic bottom-up and top-down processes involved in group affect [10], [19], and (ii) addressing the difficulty of conceptualizing group affect within complex purposive group interaction settings [9], [37], [41]. To achieve this, we recruited students with an organizational psychology background, either at the bachelor's or master's level. They were familiar with the topic of affect in groups through their study curriculum. A total of eight annotators (3 male and 5 female; ages 18-25) were recruited.

3.3 Initial Pilot Studies

We began with an initial pilot study to (1) refine the annotation procedure for the social construct and dataset, and (2) train our annotators. We selected three videos from the MeMo corpus that were perceived to have the maximum

2. Link to be available after the review process.

affective variances, which were then randomly assigned to four out of the eight annotators to annotate group-level arousal and valence using INTERACT [49]. As a primer for further training of annotators, we held individual meetings with each of the four annotators to discuss the definitions of dynamic group affect and the circumplex model.

3.3.1 Tuning the Annotation Procedure

We tuned the annotation procedure in terms of two key parameters: (i) the *scale* to be used to annotate group affect and (ii) the *time-window size* to be used to aptly capture the affective dynamics. For the pilot study, we began by utilizing the Self Assessment Manikin (SAM) [50] to annotate group-level arousal and valence [16], employing 20-second time windows in accordance with [37].

After the four annotators completed the annotations of the videos part of the pilot study, a meeting was setup with the annotators to discuss on the two parameters to be tuned. The discussion during the meeting was specifically on two questions: (i) "Should we increase or decrease or keep the same time-window size?", and, (ii) "Do you accept the scale used? Does it well capture the affective expressions and fluctuations in the interactions?".

Based on the consensus reached during this discussion, we made two changes to the initial setup for the annotation procedure. First, we reduced the time-window size to 15 secs. This decision was based on the consensus that the dynamics of group affect in the videos fluctuate rather faster and a smaller time-window size would capture the fluctuations more adequately. Second, we replaced the SAM scale with an ordinal scale ranging from 1 to 9. All annotators felt that the SAM scale had extreme arousal and valence categories on the scale that were not usually present in the observed spontaneous interactions in the MEMO corpus. Because of this, the resulting affect annotations had only limited affective variance. Our decision to replace the SAM scale with an ordinal scale aligns with arguments in favor of the ordinal nature of emotions in previous work [51].

We used an evidence-based approach to determine the most appropriate window-size to capture meaningful fluctuations in observed group-level affect. To this end, we iterated our pilot study until a consensus was reached between annotators to freeze the annotation procedure. While the annotators in our pilot study agreed that the initial window size of 20 secs was not adequate to capture all fluctuations in collective group affect, they reached a consensus that a 15 secs window was more effective in capturing these fluctuations. They also concurred that a smaller window size was unnecessary and would not add value in terms of capturing additional nuance in group affect dynamics. This choice aligns with the literature which indicates that collective group-level affect tends to change somewhat more slowly compared to individual-level affect [7]. More specifically, the development of group affect necessitates the expression of individual affect and the occurrence of a bottom-up process, leading to relatively slower fluctuations of collective group affect compared to fluctuations in individual affect.

Moreover, during the iterative tuning rounds, a critical problem raised by most of the annotators was that a particular moderator in the MeMo corpus dominated most of the interaction preventing the interaction from being an active discussion amongst all group members. With respect to this, we got rid of all 9 interactions from that particular moderator. After this, the final dataset had in total 35 group interactions. We attribute the richness of the discussions with the annotators to their educational backgrounds and the prior experience of some in annotating social behaviors.

3.3.2 Video Markers

A pitfall in using an ordinal scale over the SAM scale is that the annotators do not have a reference affective expression associated with each ordinal value of the scale, such as the illustrations in SAM. This leaves room for interpretation and confusion, as annotators are unsure when to assign a relatively higher or lower value on a scale. To tackle this problem, when training our annotators, we used *video markers* developed for each point on the ordinal scale. The objective was to link each point on the scale to a specific expression of collective group affect by identifying a fitting 15 secs video segment from within the MeMo corpus, which would serve as a behavioral anchor for the respective point on the scale. This video marker-based training of the annotators was inspired by the nonverbal behavioral anchors for affect expressions developed by Bartel and Saavedra [27], which were subsequently adopted by other works for the annotation of group affect and group mood (e.g., [10], [36], [38]). Notably, whereas Bartel and Saavedra [27] devised a list of nonverbal behaviors describing each point of their rating instrument, we aimed to contextualize this information and provide a temporal setting for each type of group affect expression, so annotators would have specific video segments that clearly illustrated each point on our ordinal scale for annotating group affect, for their reference during the annotation procedure.

To create these video markers, we began by utilizing the annotations derived from the fine-tuning processes of the annotation procedure. We selected several potential candidates for each element of the ordinal scale. These candidates were discussed with the four annotators who participated in the pilot study, using the two definitions provided to them (i.e., the definition of dynamic group affect and the circumplex model). Of note, the development of video markers was carried out alongside the training of the annotators. Throughout the training and the accompanying discussions among the annotators, the video markers were consistently adjusted, with the goal of establishing a single, definitive video marker for each element of the ordinal scale. The specific topics covered in these discussions are presented in the next section.

3.3.3 Training Annotators

The main objective of the training was to help annotators focus on the most important aspects of dynamic group affect, with discussions conducted on the following topics.

Aggregation of affective expressions : Two key ideas were discussed under this topic: (i) "How do you *aggregate the individual affective* expressions to group affect?" (i.e., on the bottom-up phenomena), and, (ii) "How do you track and *aggregate the dynamic fluctuations* in group affect?" (i.e., on the top-down phenomena). Moreover, owing to the ordinal nature of scale, the scale differences between video markers belonging to adjacent scale elements were also discussed.

	Ours		Comparison	
	Arousal	Valence	Arousal	Valence
Quadratic κ	0.41	**0.58**	**0.50**	0.54 [18]
Cronbach's α	**0.82**	**0.89**	0.80	**0.89** [37]
Correlation (PCC)	**0.51**	**0.64**	0.44	0.41 [16]

TABLE 1: Inter-annotator agreements. Note that the values in the comparison column are from individual affect studies.

	Arousal		Valence	
All	0.407		0.578	
Excluded Annotator	κ	Δ	κ	Δ
1	0.402	-0.005	0.573	-0.005
2	0.398	-0.009	0.572	-0.006
3	0.416	$+0.009$	0.581	$+0.003$
4	0.402	-0.005	0.575	-0.002
5	0.412	$+0.005$	0.567	-0.011
6	0.401	-0.006	0.598	$+0.020$
7	0.399	-0.008	0.569	-0.009
8	0.385	-0.022	0.582	-0.004

TABLE 2: Quadratic weighted kappa κ scores when a particular annotator is excluded. Δ is the increase $(+)$ or decrease $(-)$ in κ when the annotator is excluded.

Contribution of the Group Facilitators: The role and the contribution of the group facilitator in the interaction towards the group affective state was also discussed. To capture the unscripted nature of the interaction, we instructed the annotators to treat the moderator as one of the group members and not to perceive them differently from the participants, especially in light of the bottom-up nature of collective group affect.

Focusing on Affective Expressions: Research on emotional contagion and group affect indicates that behaviorally expressing an emotion, such as smiling, can lead to experiencing it, such as feeling happiness [38]. Accordingly, in this work, we instructed annotators to focus solely on the affect expressed by group members, following the circumplex model of affective expressions [35]. For example, when one of the observed group members used sarcasm (i.e., a positive valence expression to convey negatively valenced conversational content), the annotators were instructed to only focus on the affect *expressed* (in this case, positive valence). This approach generally aligns with the affect recognition literature that prioritizes *expressed* over experienced emotion [14], [16], [17].

Based on these discussion topics, one candidate was selected as the emotion marker for a particular scale item. These emotion markers are then used as reference videos to explain the respective scale item. At the end, all other annotators who were not part of the initial pilot studies were also trained using the outcome of the studies and the video markers derived for the ordinal scale.

3.4 Annotation Procedure

3.4.1 Annotation Software Setup and Location

During training and for the entire annotation procedure that followed, we used INTERACT software [49] which provides a graphical user interface where annotators can scroll through videos to annotate observed behavior directly from an audio or video file. The software allowed us to systematize and synchronize the entire annotation procedure across annotators. INTERACT was set up to request an annotation for every 15 seconds segment of each video. Clicking on each segment would play the respective slice of the video. The annotators were allowed to watch each 15 seconds time-window any number of times. They used their number pad in the keyboard to input a number between 1 and 9 as their ordinal scale annotations, and any wrong input would not be accepted by INTERACT. To ensure an appropriate setting without distractions, annotators were asked to come to the laboratory where they were provided with an individual workspace, including a desktop computer with INTERACT installed and a two-screen setup.

3.4.2 Distribution of Videos

The 35 interaction videos were provided in a randomized order to each of the annotators, to prevent any potential annotator bias which might occur if all the annotators received the videos in the same order. The videos were distributed in batches of 5 videos to ensure that the annotators followed the order provided. The annotators worked 10-20 hours per week, annotating approximately 2-5 videos per week. The annotation procedure ensured that each group discussion received at least five annotations throughout (i.e., five annotations for each 15-second window of each of the 35 group discussions). In comparison, most state-of-the-art individual-level affect recognition datasets have three to five annotators [17], [52], with literature on uncertainty modeling suggesting that at least four annotations should be collected for a reliable annotation distribution and its ground-truth consensus [53]. The annotations for arousal and valence were done independently, similar to [17], [18], with the annotators watching the entire video separately while annotating for each dimension of affect.

3.4.3 Inter-Annotator Agreement

To measure the inter-annotator agreement, we used three different metrics: (i) quadratic weighted kappa (κ) [54], (ii) Cronbach's alpha (α) [55], and (iii) Pearson's correlation coefficients (PCC) [56]. The three metrics are chosen with their respective advantages in mind. Firstly, the quadratic weighted kappa measure κ is a variant of Cohen's kappa that is specifically designed to measure agreements in the ordinal scale data [54]. However, it is constraint to calculating agreement between only pairs of annotators. Hence, following common practices [18], [26], [48], we report the average in all possible combinations of annotators. Secondly, Cronbach's alpha (α) measure overcomes this limitation by allowing one to measure the agreement between an arbitrary number of annotators. Finally, as we annotate a time-dependent dynamic construct, we also use the PCC measure as the inter-annotator agreement metric.

As the group affect annotations collected here are novel, both in terms of the social setting at hand and its dynamic nature that captures the temporal fluctuations, there is no direct comparison that can be made in terms of the agreement scores. However, to better understand the agreement scores, we compare their value with other agreement scores reported in the existing literature on collecting affect annotations (see Table 1). The criteria for selecting the litera-

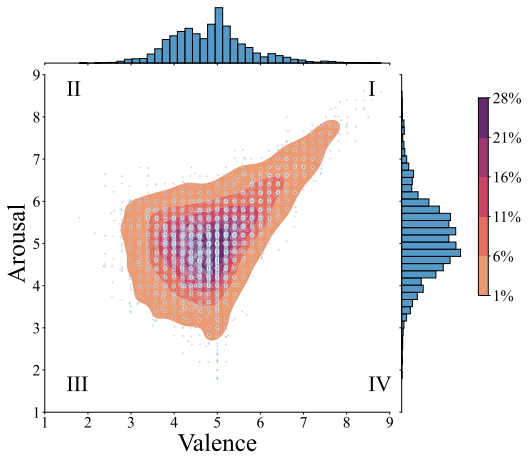


Fig. 1: Distribution of final ground-truth for dynamic group affect $y_{s,t}^{\text{EWE},(i)}$, in terms of arousal and valence.

ture to compare with is as follows: (i) the annotations are performed on affect either at the individual-level or group-level, (ii) the agreement scores are presented on the metrics chosen above. Note that no hard filter was set on the social setting at hand, due to the less availability of literature on group affect in an purposive interaction setting.

Quadratic weighted kappa κ measure : From Table 1, we observe a κ =0.41 and κ =0.58 for arousal and valence, respectively. This indicates a moderate agreement for both arousal and valence dimensions of group affect [54]. As a comparison, which is also presented in Table 1, the MSP-Conversation dataset [18], on a much simpler annotation task of individual-level affect, has an agreement of κ =0.50 and κ =0.54 for arousal and valence, respectively (i.e., higher than ours on arousal and lower on valence).

Cronbach's α measure : From Table 1, we observe an α of 0.82 and 0.89 for arousal and valence, respectively, indicating a good and satisfactory level of agreement [55]. For comparison, the annotations of individual affect in groups collected by [37] have an α agreement of 0.80 for arousal and 0.89 for valence, which is lower than our arousal annotations and virtually the same as our valence annotations. Note here that [37] worked on a much simpler social settings without accounting for interactions between interlocutors.

Pearson Correlation (PCC) measure : From Table 1, we observe a PCC of 0.51 and 0.64 for arousal and valence, respectively. The individual-level affect annotations in the RECOLA dataset [16], in comparison, have a PCC agreement of 0.435 for arousal and 0.407 for valence, which is much less than our agreement scores, despite being on a much more simpler social construct.

From Table 1 we note that in general the agreement for valence is higher than that of arousal, which aligns with the literature [18], [53]. Furthermore, in Table 2, we present the κ scores when a particular annotator was excluded, where Δ is the increase $(+)$ or decrease $(-)$ in κ when the annotator was excluded. Table 2 reveals that in most cases (i.e., for six out of eight annotators), excluding the annotator only decreases the κ score $(-)$ in terms of both arousal and valence. Only for annotator 5 a large increase in Δ is noted when excluded, i.e., Δ of $+0.020$ is noted for valence. In all other cases, the Δ is rather minimal, i.e., $\Delta < +0.01$. This points to the reliability of the collected annotations, even when annotations from an annotator are omitted.

3.5 Ground-truth

To derive the final ground-truth of dynamic group affect, we use the Evaluator Weighted Estimator (EWE) [57], a common technique to derive ground-truth for individual-level affect recognition [16], [18]. The EWE, to derive the ground-truth, weights annotations with respect to the annotator-wise correlation coefficients, and is formulated for a time segment t in an interaction sample s as follows:

$$y_{s,t}^{\text{EWE},(i)} = \frac{1}{\sum_{k=1}^{K} r_{k,s}^{(i)}} \sum_{k=1}^{K} r_{k,s}^{(i)} \quad y_{s,t}^{(i)} \tag{1}$$

where $r_{k,s}^{(i)}$ denotes the average correlation coefficient of the annotator k with other annotators for a particular interaction sample s and emotion dimension i. For unreliable annotators with $r_{k,s}^{(i)} < 0$ a lower bound of zero is defined. In cases where all k annotators have the same correlation coefficients $r_{k,s}^{(i)}$, they result in the same correlation weighting and thereby $y_{s,t}^{\text{EWE},(i)} = y_{s,t}^{(i)}$. For the rest of the experiments and analyses in this work we will use $y_{s,t}^{\text{EWE},(i)}$ as the ground-truth of dynamic group affect.

The distribution of the ground-truth in terms of the arousal and valence dimensions can be seen in Figure 1. From the plot, we note that the independently annotated arousal and valence, depicted by the histograms of the marginal distribution, have considerable variances. Arousal annotations are distributed with a mean (m) of 5.10, a standard-deviation (s) of 0.88, a minimum value $(\min)$ of 1.80, and a maximum value $(\max)$ of 8.61. Similarly, the valence annotations are distributed with $m = 4.84$, $s = 0.90$, $\min = 1.80$, and $\max = 8.80$.

Samples in Quadrant I denote high-arousal and positive-valence (e.g., emotions such as happy and excitement). Quadrant II denotes high-arousal and negative-valence (e.g., emotions such as anger and frustration). Quadrant III denotes low-arousal and negative-valence (e.g., emotions such as depressed and gloomy). Quadrant IV denotes low-arousal and positive-valence (e.g., emotions such as relaxed). With respect to the joint distribution of arousal and valence, depicted by the heatmap and the scatter plots (in Fig.1), the variances are noted only in some of these quadrants of the circumplex model. For example, good number of samples are noted in Quadrant I. Similarly, a good number of samples can also be noted for low-arousal and neutral-valence (i.e., between Quadrant III and IV; emotions like tired and melancholic), and for neutral-arousal and negative-valence (i.e., between Quadrant II and III; emotions like bored and sad). However, in Quadrant IV and Quadrant II extreme samples are not noted. This could be because the participants in MeMo were reported to be non-preacquainted and complete strangers at the beginning of the longitudinal study that spanned over 3 interactions in two weeks. In such short longitudinal cases, amongst non-preacquainted participants, extreme expressions are very rare [58], likely due to professional behavioral norms.

See Appendix Section 2 for examples of the collected annotations and corresponding video frames, providing a qualitative illustration of how the annotations capture the dynamic ebb and flow of group affect. These examples, together with the inter-annotator agreement results, thereby substantiate the *first contribution*—the development of an annotation strategy for group affect that establishes theory-method alignment with concepts from organizational psychology—outlined in Section 1.

4 Modeling Group Affect Dynamics

This section outlines our approach to modeling dynamic group affect. We begin with preprocessing of raw audio and video data (Section 4.1), followed by the extraction of individual-level features (Section 4.2). Finally, in Section 4.3, we present two complementary methods for modeling group affect as a dynamic and collective construct: (i) handcrafted features capturing synchrony and convergence, and (ii) a graph neural network-based approach that models interpersonal relationships in a data-driven manner.

4.1 Preprocessing

The MeMo corpus provides with manually diarized and synchronized audio for each of the interlocutors, collected at a sample rate of 16kHz [47]. Similarly, video recordings of online group discussions are also provided at a frame rate of 60 fps. The group discussion video frames are cropped to obtain individual-level frames of each of the interlocutors.

4.2 Individual-level Feature Extraction

Existing research works have revealed the multimodal nature of affect. For instance, the audio modality has shown to be more informative of the arousal dimension of affect, whereas the video and text modalities better explain the valence dimension [59]–[62]. To this, we employ a multimodal strategy to model group affect, by employing both audio-based and video-based features.

Audio features : As the audio features we extract the first 5 *MFCC* coefficients, *Voice Intensity*, *Pitch*, *VGGish*, and the *Speech Rate*. This set of individual-level paralinguistic audio cues are selected owing to their demonstrated effectiveness in the automatic recognition of individual-level affect [63] and also other group-level constructs such as cohesion [31]. The MFCC and pitch features were calculated using the `librosa` package, for every 10 ms with a sliding window of 30 ms. The voice intensity and speech rate features were extracted using `Praat` [64]. The voice intensity was calculated at the same rate as the MFCCs while the speech rate was calculated in vowels per second using [65] at a rate of 1.5 secs following [31]. The VGGish features are pretrained deep learning based features and was extracted using the pretrained weights from [66].

Video features : As the video features we extract *Facial Action Units (AUs)*, *Face Pose*, and *ResNet50*. Action units, Face pose and ResNet50 features have been successfully used in existing literature for several tasks, such as sentiment analysis [67] and emotion recognition [29]. Individual-level AUs (subset available in the `OpenFace` toolkit) and the face pose (pitch, roll and yaw) were extracted using the `OpenFace` toolkit [29], for every 0.5 sec. Similar to the VGGish audio features, the ResNet50 network was used to extract framewise pretrained deep learning based features and was extracted using the pretrained weights from [68].

4.3 Group-level Modeling Techniques

4.3.1 Handcrafted features based group modeling

Social interactions are multilevel systems where interpersonal relationships and affective states emerge at multiple levels of the interaction, i.e., at the individual, dyadic and group level [69]. With respect to this theoretical framework of group-level constructs, in this work, to study dynamic group affect, from the individual-level features we extract dyadic-level and group-level features that are descriptive of the interpersonal relationships shared between a dyad in the interaction and the group as a whole, respectively. The complete list of handcrafted features extracted can be seen in the Appendix Table S1.

Dyad-level features: We extract two sets of dyad-level interpersonal relationship-based features: (1) Synchrony and (2) Convergence. As the synchrony feature set, following [26], we use linear correlation coefficient-based measures: (i) the correlation coefficient ρ, the linear correlation without a time-lag, (ii) lagged correlation ρ_δ, the linear correlation with the best time-lag, and (iii) the best lag δ defined as the time-lag used to obtain the maximum linear correlation between the two individual-level signals. Existing literature notes that synchronous behavior is displayed by interlocutors often in a time-lagged manner, with a leader and a follower [26], [30]. The three synchrony measures are extracted using the formulation below:

$$\begin{aligned} \text{Correlation coeff. } \rho &: X \circledast Y \\ \text{Lagged correlation } \rho_\delta &: \max z(X,Y) \\ \text{Best lag } \delta &: \arg\max_l z(X,Y,l) - ||X|| + 1 \end{aligned} \tag{2}$$

$$z(X,Y,l) = \sum_{k=0}^{||X||-1} X_l \circledast Y_{k-l+N-1} \tag{3}$$

where z(., .) is the cross-correlation function, $\circledast$ denotes linear correlation between two signals, $||X||$ denotes the length of signal X, $l = 0, 1, ..., ||X|| + ||Y|| - 2$ denoting the time-lags possible, and $N = \max(||X||, ||Y||)$.

Following the technique proposed in [70] and [26], we extract three convergence features: (i) global, (ii) symmetric and (iii) asymmetric convergence. Global convergence captures the change in similarity between two individual's social signals, specifically between the initial time-segments and the later time-segments. Similarly, symmetric and asymmetric convergence features capture the decrease or increase in similarity between the two individual's social signals, without and with a time-lag, respectively. The three measures are formulated as:

$$\begin{aligned} \text{Global } \Theta_{\text{gbl}}&: \sum_{i=0}^{||X||/2} (X_i - Y_i)^2 - \sum_{j=||X||/2}^{||X||} (X_j - Y_j)^2 \\ \text{Symmetric } \Theta_{\text{s}}&: (X_l - Y_l)^2 \circledast L \\ \text{Asymmetric } \Theta_{\text{as}}&: p(Y_b/\theta_{X_a}) \circledast L \end{aligned} \tag{4}$$

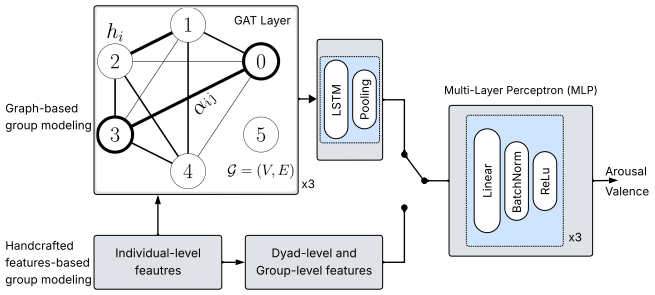


Fig. 2: Overview of the predictive modeling architecture.

where, $L = [0, 1, ..., ||X||]$, $l \in L$, and θ is the parameter of a Gaussian mixture model (GMM) trained using the expectation-maximization procedure on the data points from X in the initial period of the interaction, i.e., $a \in [0, m], m = 2 \cdot ||X||/3$. Similarly, Y_b are data points from Y in the later period of the interaction, i.e., $a \in [m, ||Y||]$.

Group-level features: To extract *group-level* features from the individual and dyadic features, following [26], [31], we use six aggregation techniques that are agnostic to group size: `average`, `standard-deviation`, `median`, `minimum`, `maximum`, and `gradient`. The core idea here is that these different types of aggregations, each with an unique approach, describe the distribution of dyadic-level features within a group thereby capturing the interpersonal nuances within all possible dyads in the group. The `average` and `standard-deviation` explains the average and the deviation from the average of synchrony measures across all possible dyads in a group. Similarly, the `gradient` aggregator explains the deviation between the least and most synchronous dyad, i.e., the absolute difference between the `minimum` and `maximum`.

It is important to note that the group-level aggregations were computed from fine-grained temporal features at the dyad level, capturing detailed nuances in the temporal dimension. However, one could argue that these aggregations may miss finer details in the interpersonal relationship dimension. To address this, we propose using GNNs in the following section as a data-driven alternative to handcrafted features for modeling group affect.

4.3.2 Graph-based group modeling

Grounded in social network theory [32], we frame group affect modeling as a graph classification task using GNNs, where the group interaction is represented as an undirected graph $\mathcal{G} = (V, E)$, with V denoting the set of M nodes and E the set of edges. Each node $V_i \in V$ represents an interlocutor with individual-level features $h_i \in \mathbb{R}^F$, and each edge $E_{i,j} \in E$ denotes a connection between two nodes. The adjacency matrix A captures the edge structure, where $A_{i,j} = 1$ if nodes i and j are connected and $A_{i,j} = 0$ otherwise.

A standard GNN training pipeline, as introduced in Graph Convolutional Networks (GCN) by Kipf et al. [71], consists of two main steps: (1) *convolution*, where each node transforms its feature representation (h_i) to share with adjacent nodes, and (2) *message passing*, wherein these features are propagated to the adjacent nodes. Nodes subsequently update their representations by aggregating information from their neighbors, typically using simple operations such as summation or averaging. However, this approach can produce identical output features for nodes with identical neighborhoods, thereby limiting the model's expressiveness for certain graph structures. For example, this poses a challenge in group affect modeling, where interlocutors are connected through a fully connected graph that captures overall group membership but overlooks the distinct relationships between individuals. To address this limitation, we employ an attention-based message passing mechanism, as proposed in Graph Attention Networks (GAT) [72].

Unlike basic sum or average aggregations in GCN, the attention based GAT compute a weighted average of multiple node features. These weights are dynamically determined based on a combination of the node's own features, the interlocutor's individual-level features (h_i), and the features of adjacent nodes (h_j), which represent the interacting counterparts of the interlocutor. The GAT layer obtains an attention based aggregation formulated as:

$$h_i^{(l+1)} = \phi\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}^{(l)} h_j^{(l)}\right), \tag{5}$$

where $\mathbf{W}^{(l)}$ is the learnable weight parameters that transforms the node features, ϕ represents an arbitrary activation function, and α_{ij}, the learnable attention weight between nodes i and j. The attention weight α_{ij} is formulated as:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}\left[\mathbf{W}h_i || \mathbf{W}h_j\right]\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}\left[\mathbf{W}h_i || \mathbf{W}h_k\right]\right)\right)}, \tag{6}$$

where $||$ denotes concatenation, $\vec{\mathbf{a}}$ represents the attention mechanism, implemented as a single-layer feedforward neural network parameterized by a weight vector $\vec{\mathbf{a}} \in \mathbb{R}^{2F}$, LeakyReLU is the chosen activation function, and $\mathcal{N}_i$ denotes the indices of the adjacent nodes of node i. To accommodate varying group sizes (g), we fix the number of nodes M to the maximum group size in the dataset. For groups with fewer than M members, dummy nodes without edges are added, ensuring that $A_{x,j} = 0$ for such nodes ($g < x \leq M$).

The overall architecture consists of three GAT layers followed by an LSTM, enabling temporal modeling across consecutive 15-second time segments. Mini-batching is employed to ensure that each batch contains segments from the same interaction and maintains temporal continuity. This batching strategy is commonly used in end-to-end continuous emotion recognition techniques [53], [59], [73].

5 Results

This section analyzes and discusses the results of group affect modeling. First, Section 5.1 presents predictive modeling based on the collected group affect annotations. Then, Section 5.2 explores quantitative analyses of the convergence and divergence phenomena.

5.1 Predictive Modeling

Using the handcrafted features described in Sec. 4.3.1 and the graph neural network in Sec. 4.3.2, we frame the predictive modeling of group affect as a regression task. Figure 2

	Feature Set	TM	Model	Arousal CCC ↑	Valence CCC ↑	Avg. CCC ↑
Audio	Basic	×	*HCF*	0.242	0.245	0.244
	Sync.	×	*HCF*	0.215	0.203	0.209
	Comb.	×	*HCF*	0.261	0.222	0.241
	Comb.	✓	*HCF*	0.274	0.242	0.258
	Indiv.	×	*GAT*	0.253	**0.263**	0.253
	Indiv.	✓	*GAT*	**0.286**	0.255	**0.271**
Video	Basic	×	*HCF*	0.315	0.405	0.360
	Sync.	×	*HCF*	0.342	0.428	0.385
	Comb.	×	*HCF*	0.355	0.448	0.401
	Comb.	✓	*HCF*	0.339	0.445	0.392
	Indiv.	×	*GAT*	0.344	0.456	0.400
	Indiv.	✓	*GAT*	**0.361**	**0.483**	**0.422**
Audio-Visual	Basic	×	*HCF*	0.293	0.332	0.313
	Sync.	×	*HCF*	0.403	0.428	0.401
	Comb.	×	*HCF*	0.416	0.431	0.424
	Comb.	✓	*HCF*	0.396	0.447	0.422
	Indiv.	×	*GAT*	0.384	0.484	0.434
	Indiv.	✓	*GAT*	**0.420**	**0.508**	**0.464**

TABLE 3: Results of the predictive modeling. TM: Temporal Modeling, *HCF*: Handcrafted features, *GAT*: Graph Attention Network.

presents the block diagram of the modeling technique's architecture. As shown, a simple Multi-Layer Perceptron (MLP) performs the regression task, using as input either the handcrafted features or the average pooled output of the GAT layers. The MLP is made up of three linear layers with ReLU and Batch Norm after each of the layers. The MLP's architecture was tuned with respect to the loss obtained on the validation dataset. The code for the modeling technique, including the extraction of handcrafted features and the implementation of the graph neural network, is publicly available and can be found here[3].

5.1.1 Experimental Setup

Feature Sets: To evaluate the predictive capabilities of the extracted *handcrafted features* (**HCF**), we employ three sets of group-level features: (1) Basic, where individual-level features are directly aggregated to the group level using group-level aggregators. This set does not account for interpersonal or dyadic relationships and only partially captures social signal dynamics due to the use of average temporal and group-level aggregations. (2) Synchrony (or Sync.), where dyadic synchrony and convergence-based features are extracted before applying group-level aggregations. This set effectively captures both social signal dynamics and interpersonal relationships among interlocutors. (3) Combined (or Comb.), which fuses the Basic and Synchrony feature sets, integrating both aspects into a unified model. These feature sets are further classified into audio, video, and audio-visual categories.

Contrarily, to study the predictive capabilities of the proposed *graph based modeling* (**GAT**), we use only individual-level features (or Indiv.) as input, categorized into audio, video, and audio-visual feature sets. For this, we initialise the node features h_i of the graph $\mathcal{G}$ with the individual-level features of the respective interlocutor.

Loss Function: The concordance correlation coefficient (CCC) [74] is used as the loss function and as the metric to validate the performances. The CCC has been widely used in literature for the task of individual-level affect recognition [14]. The CCC measures the agreement between two variables and ranges from -1 to $+1$, with perfect agreement at $+1$. In contrast to Pearson's correlation, the CCC takes both the linear correlation and the bias in to account, which makes it preferable over Pearson's correlation as the loss function and as the evaluation metric.

Training Strategy: The models are trained using the ADAM optimizer with a learning rate of 10^{-4}. The strategy includes an early-stopping on the validation loss improvements with a patience of 10 epochs. The best model during training is selected as the one with the best validation loss.

Data Partition: The dataset is partitioned following the strategy proposed in [17]. The partitions were made such that there is no speaker overlap between the training and testing datasets. This also includes non-overlapping moderators in the MeMo corpus. The validation dataset however may have an overlap of moderators in some samples, but not overlapping participants. Overall, the training-testing split is made with a 80-20% split, and 10% of the training datasets is split for the validation dataset.

5.1.2 Discussion

Table 3 presents the results of the predictive modeling. The performance is evaluated across four feature sets (Basic, Synchrony, Combined, and Individual-level), three modalities (Audio, Video, and Audio-Visual), two modeling techniques (HCF and GAT), and two temporal settings (with and without modeling temporal relationships between consecutive 15 second segments).

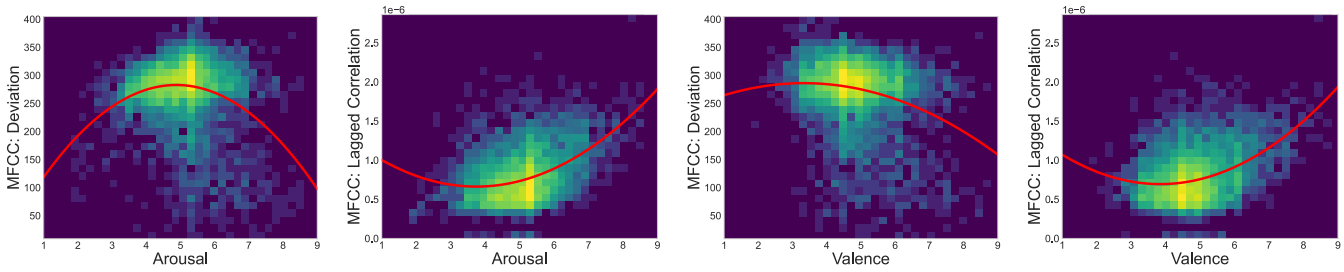
Multimodal nature of group affect: The results reveal that dynamic group affect is best captured in a multimodal manner, with the audio-visual feature set obtaining the best performance of 0.420 and 0.508 in terms of arousal and valence, respectively. Furthermore, the visual feature set outperforms the audio feature sets in predicting group affect, demonstrating superior performance across feature sets and modeling techniques. We also note that the audio modality better explains the arousal dimension than the valence dimension, while the video modality better predicts the valence dimension. However, their performance differences are not as large as noted in individual-level affect recognition literature [62].

Relevance of temporal modeling: In Table 3, a ✓ in the "TM" column indicates that the method models temporal relationships between consecutive 15-second segments, while a × denotes it does not. Capturing these temporal dynamics allows the model to represent affective processes such as the top-down influence of group-level affect on individuals through emotional contagion (as discussed in [2], [75], and partially explored in [19]). Results show that temporal modeling notably enhances group affect prediction, particularly in GAT-based models, where CCC improvements are more consistent across modalities. To our knowledge, this is the first work to effectively capture this top-down mechanism of group affect, addressing a key limitation in prior studies [9], [19] (as discussed in Sec. 2.3). This was made possible by collecting group affect annotations that reflect temporal dynamics.

3. https://github.com/sp-uhh/group_affect

		Arousal					Valence				
		Regression Analysis				Kendal's	Regression Analysis				Kendal's
		α	β	c	R^2	τ	α	β	c	R^2	τ
Pitch	σ	-2.038	+22.322	-12.183	1.2%	−0.081	-0.650	-6.654	+63.554	1%	−0.002
	ρ	+0.002	-0.019	+0.023	6%	+0.210*	+0.002	-0.013	+0.001	4%	+0.212*
VGGish	σ	-0.090	+0.851	+0.646	2.1%	−0.105	-0.027	-0.462	+4.168	3.7%	−0.194*
	ρ	+0.013	-0.146	+ 1.123	7.2%	+0.027	-0.005	+ 0.018	+ 0.660	2%	+0.171*
	Θ_{as}	+0.030	-0.026	+0.324	13%	+0.112	-0.001	+0.018	+0.211	18%	+0.168*
MFCC (1st Coeff.)	σ	-10.925	+106.629	+22.751	7.7%	−0.170*	-3.957	+26.370	+242.151	6.4%	−0.210*
	ρ_δ	+0.010	+0.029	+0.001	16.4%	+0.293*	+0.021	+0.000	+0.017	15%	+0.270*
	Θ_s	+0.001	+0.008	-0.022	2%	+0.023	+0.003	-0.034	+0.094	2%	+0.037
AU06	σ	-0.002	-0.010	+0.058	3.3%	−0.210*	-0.001	-0.004	+0.077	6.6%	−0.259*
	ρ	+0.018	-0.125	+0.209	15.5%	+0.264*	+0.017	-0.102	+0.150	18.5%	+0.275*
AU07	σ	+0.002	+ 0.001	+0.067	7.0%	−0.213*	-0.002	0.025	-0.004	14.8%	−0.343*
	ρ	+0.077	-0.059	+0.107	2.0%	+0.138	+0.067	-0.047	+0.076	2.0%	+0.141
AU12	σ	-0.004	-0.007	+0.277	1.5%	−0.104	-0.003	+0.02	+0.211	2.2%	−0.130
	ρ	+0.019	-0.135	+0.240	16.1%	+0.230*	+0.015	-0.085	0.115	18.4%	+0.272*
AU25	σ	-0.012	+0.009	+0.135	3.1%	+0.027	-0.004	+0.069	-0.118	3.5%	+0.233*
	ρ	+0.013	-0.109	+0.224	8%	+0.125	+0.012	-0.090	+0.172	8%	+0.132
Head Roll	ρ_δ	+0.178	-1.799	+14.633	2%	+0.015	+0.113	-1.243	13.511	2%	+0.035

TABLE 4: Quantitative analysis of the convergence-divergence processes.



(a) Standard deviation (b) Time-lagged correlation (c) Standard deviation (d) Time-lagged correlation

Fig. 3: Relationship between convergence-divergence measures and group affect: *Arousal* (a,b) and *Valence* (c,d).

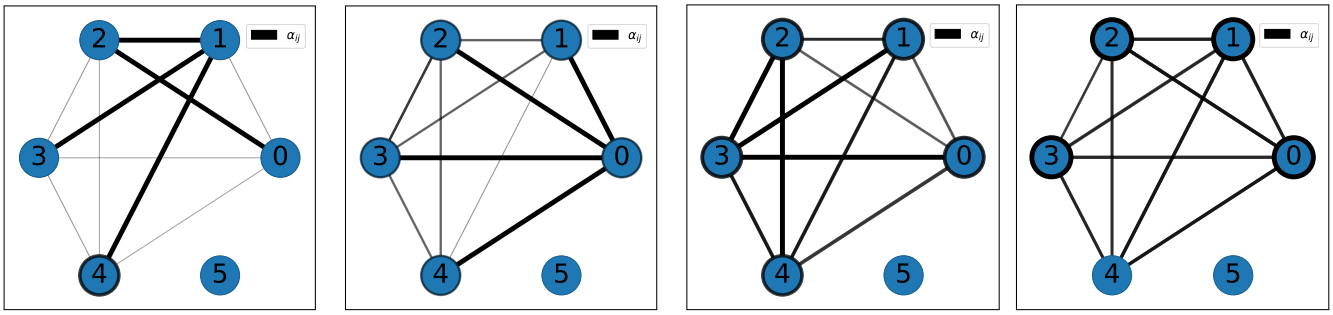
Relevance of capturing interpersonal dynamics: The results highlight the significance of capturing interpersonal dynamics in group affect modeling, as both HCF and GAT modeling techniques demonstrate improved performance. For example, incorporating synchrony and convergence features into the video and audio-visual sets increases average CCC from 0.360 to 0.401 and from 0.313 to 0.445, respectively. However, in the audio modality, synchrony features perform worse than the basic set, likely due to segments where most interlocutors are silent, limiting synchrony extraction. This issue does not occur in the video modality, where participants are typically active, such as in attentive listening scenarios [76].

HCF vs GAT for capturing interpersonal relationships: Regarding the modeling technique used to capture interpersonal relationships, the data-driven GAT consistently outperforms the handcrafted synchrony and convergence features (HCF) across all modalities. The most notable improvement is observed in the audio-visual setting, where GAT achieves a CCC of 0.410 for arousal and 0.508 for valence, compared to HCF's 0.396 and 0.447, respectively. The performance gain is greater for valence than for arousal, highlighting GAT's stronger ability to model interpersonal dynamics associated with emotional valence. The GAT model's improved performance underscores its effectiveness in modeling group affect by capturing micro-level, multimodal social dynamics.

5.2 Analysis on Affective Convergence and Divergence

With the majority of empirical research focused on static group affect [9], [19], [22], Kelly & Barsade [75] emphasized on the dynamic nature of affect, i.e., how, over time, the nature of collective affect can change. The ebb-and-flow of collective group affect over time is primarily characterized by the affective convergence and divergence underlying the bottom-up and top-down processes of group affect [2], [28]. Foundational theories on affective dynamics (e.g., [45] and [28]) further describe how several individual interaction- and behavior-level mechanisms, including facial mimicry, emotional similarity and dissimilarity, and empathy, contribute to affective *convergence* and *divergence* in groups.

Building on these theorizations, in this section, we present a quantitative analysis on the relationship between interaction- and behavior-level cues, that quantify the level of affective convergence and divergence within interlocutors, and the collected annotations of dynamic group affect. To this end, we analyse affective convergence and diver-



(a) Arousal:0.75, Valence:0.73 (b) Arousal:0.29, Valence:0.35 (c) and (d): Arousal:0.47, Valence:0.50

Fig. 4: Visualization of attention weights α_{ij}, for Positive (a), Negative (b), and Neutral (c and d) samples.

gence using the two techniques, HCF and GAT, that were employed in the predictive modeling discussed in Sec. 5.1.

5.2.1 Experimental Setup

Analyses using HCF: As the *independent variable*, we use the group-level handcrafted features that explain the within-group convergence and divergence, the similarity and dissimilarity in social signals amongst interlocutors. Specifically, the `mean` (μ) aggregation of the *dyad*-level features, and the `standard-deviation` (σ) of the *individual*-level features are used. Intuitively, *larger* σ of individual level features indicate a group that is *diverging*, while *smaller* values indicate that it is *converging*. Contrarily, *larger* μ of dyad level features imply convergence and *smaller* values that of divergence. As the *dependent variable*, the ground-truth annotations of group affect is used. To model the relationship, a least-squares based polynomial regression with *two* degrees of freedom is used, where the relationship is modeled as a 2^{nd} degree polynomial in the independent variables. The regression model is formulated as $y_t = \alpha x_t^2 + \beta x_t + c$.

For the quantitative analysis the regression coefficients of the polynomial model (α, β, and c) are analyzed for all the independent measures used. Along with the regression coefficients, the polynomial model's R-Squared (R^2) is also analyzed. Additionally, the Kendal's rank correlation coefficient τ, along with its statistical significance ascertained with a two-tailed p-value $\leq 10\%$ (denoted by $*$), is used to reveal the direction (positive or negative) of linear relationship in the ordinal scale of group affect.

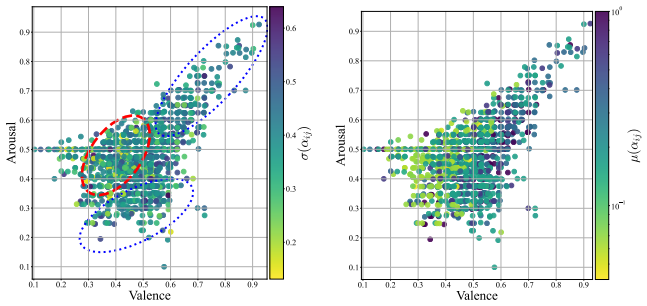
Analyses using GAT: The GAT layer models interpersonal relationships within a group by learning an attention weight α_{ij} for each dyad (as discussed in Section 4.3.2), where higher values indicate greater importance in modeling group affect—suggesting more synchronous (convergent or divergent) behavior between interlocutors i and j. Lower values imply weaker synchronization and less affective alignment. For visualization, in Figure 4 we plot the graph $\mathcal{G}$ with edge width and opacity proportional to the corresponding α_{ij} values. Additionally, we also analyze the standard deviation $\sigma(\alpha_{ij})$ and mean $\mu(\alpha_{ij})$ of attention weights of the graph at a particular time segment. To preserve the full dynamic range and expressivity of the attention weights, we compute the mean and standard deviation on the raw, non-normalized values. A lower value of $\sigma(\alpha_{ij})$ indicates that all interpersonal relationships are of equal equal (un-)importance, with none being especially prominent, whereas a higher value implies that certain interpersonal relationships are assigned a notably greater level of importance than others. Similarly, a higher value of $\mu(\alpha_{ij})$ suggests that, on average, the model assigns greater importance to interpersonal connections overall, while a lower value indicates a more uniformly low weighting across relationships.

5.2.2 Quantitative Analyses and Discussion

The convergence-divergence analyses using HCF is presented in Table 4 and Figure 3. Similarly, the analyses using GAT can be seen in Figures 4 and 5. Based on both these analyses we make the following observations.

Trends across the affect scale: We note that the interacting groups tend to *diverge* in terms of their social signals along neutral levels of arousal and valence (i.e., mid-scale values of 4 to 6) and *converge* along extreme levels of arousal and valence (i.e., strong positive affect values of 8-9, or, strong negative values of 1-2). This trend is inferred using the *negative* α values for deviation based group-level features (i.e., σ features), and *positive* α values for synchrony and convergence based features (i.e., ρ, ρ_δ, Θ_s features). Note that negative α values denote concave curves (e.g., seen in Figures 3a and 3c), and positive α values denote convex curves (e.g., seen in Figures 3b and 3d). However, there are instances where this trend does not align, such as with the ρ and Θ_{as} features of VGGish. Notably, in these cases, the degree of convexity or concavity is rather minimal; that is, the $|\alpha| \leq 0.005$. As a result, the significance of the negative or positive sign diminishes, particularly when α tends towards 0, making the relationship more linear.

Similar patterns emerge in the analysis of the attention weights α_{ij} within the GAT layers. As illustrated in Figure 4, the distribution of α_{ij} varies distinctly across different levels of arousal and valence. A notable contrast is observed between extreme affective states and neutral levels. First, in the case of extreme arousal and valence levels (see Figures 4a and 4b), several α_{ij} values stand out prominently, exhibiting significantly higher weights compared to other dyadic relationships (captured by α_{ij}) as well as individual-level contributions (captured by α_{ii}). Importantly, every interlocutor in such groups is involved in at least one high-weight dyadic connection, suggesting the presence of strong interpersonal dynamics indicative of extreme group affect levels. These groups are characterized by a *high* $\sigma(\alpha_{ij})$, reflecting greater variability in dyadic attention weights.



(a) Distribution of $\sigma(\alpha_{ij})$ (b) Distribution of $\mu(\alpha_{ij})$

Fig. 5: Distribution of the aggregates of attention weights (α_{ij}) relative to arousal and valence.

In contrast, for groups exhibiting neutral affective states, no dyadic connection is notably strong, with all α_{ij} values remain uniformly low. This indicates that interpersonal synchrony is weak or undifferentiated, and such groups exhibit a *low* $\sigma(\alpha_{ij})$. Second, for extreme affective states, dyadic interactions dominate the graph structure, with $\alpha_{ij} > \alpha_{ii}$, highlighting the significance of interpersonal relationships over individual-level cues. Conversely, in neutral conditions, individual features are equally or more influential, with $\alpha_{ij} \leq \alpha_{ii}$. Additional examples supporting these two key observations are provided in Appendix Section 5.1.

To further investigate this pattern, Figure 5a presents a plot of $\sigma(\alpha_{ij})$ as a function of the corresponding arousal and valence levels. The visualization reveals two distinct clustering patterns. First, two clusters outlined with blue dotted boundaries represent group samples exhibiting *high* $\sigma(\alpha_{ij})$ values, corresponding to *extreme* levels of arousal and valence. Second, a cluster enclosed by a red dashed boundary captures group samples with *neutral* affective states, characterized by *low* $\sigma(\alpha_{ij})$ values. Appendix Section 5.2 presents additional analyses illustrating the evolution of GAT-based attention weights during transitions between different levels of affect.

Positive vs negative affect: We note that convergence is higher for positive affect than for negative affect, in both arousal and valence, suggesting that interlocutors align their social signals more during activities that lead to the emergence of positive affect than when negative affect arises. In the context of the HCF-based analysis (see Table 4), this is reflected by the *negative* Kendall's τ for σ features, and the *positive* τ values for features related to synchrony and convergence (e.g., ρ and ρ_δ).

In the context of GAT layers, although the attention weight patterns α_{ij} appear similar for both positive and negative affect (see Figures 4a and 4b, respectively), we further examine the differences by plotting the mean attention weights $\mu(\alpha_{ij})$ as a function of arousal and valence in Figure 5b. This plot reveals that the average α_{ij} associated with group interaction segments tends to be higher for *positive* affect compared to *negative* affect. There is a clear decreasing trend in the mean attention weights, with $\mu(\alpha_{ij})$ gradually reducing as group affect shifts from high arousal-valence to low arousal-valence states. This suggests that the GAT learns attention weights α_{ij} in a way that the aggregated node representations h_i (5) positively correlate with the corresponding ground-truth affect. This behavior is consistent with the earlier observations of positive Kendall's τ coefficients (see Table 4) between group affect and handcrafted features capturing synchrony and convergence.

The results of the predictive modeling, along with the convergence-divergence analyses, thereby substantiate the *second contribution*—the multimodal modeling of dynamic group affect through capturing the underlying phenomena of convergence and divergence—outlined in Section 1.

6 Conclusions

In this work, we move beyond the traditional emphasis on individual-level affect [14], [17], [77] to address the relatively underexplored collective, group-level affect. Our contributions to this area are twofold: (1) we proposed a novel group affect annotation methodology grounded in organizational psychology theory, and (2) we introduced a multimodal modeling approach capable of capturing the complex dynamics of affective convergence and divergence underlying dynamic group affect.

For the *first contribution*, we developed an annotation strategy that not only addresses key challenges in group affect annotation, but also ensures methodological alignment with theoretical frameworks from organizational psychology [2], [58], [78]. Specifically, we tackled the critical challenge of capturing the temporal dynamics inherent in group affect. While previous literature often neglected these dynamic aspects [19], [22], [25], existing modeling techniques were also constrained by the lack of annotations that reflect the temporal context within which affective expressions occur [9], [19]. Our approach leverages an iteratively tuned 15-second window to more accurately capture the evolution and fluctuation of group affect over time. Furthermore, our study focuses on complex, purposive groups characterized by dynamic interpersonal interactions—unlike prior works [9], [19], [43], [44], which often centered on non-purposive groups lacking social intactness and goal-oriented interdependence. The inter-annotator agreement analysis on the collected annotations indicated a moderate level of agreement, as interpreted using Cohen's κ. The quality of the obtained group affect annotations is also evidenced by their ability to capture the ebb and flow of affect, including subtle differences in social signals between consecutive segments of the observed group interactions (see Appendix Section 2 for more detail).

Regarding the *second contribution*, leveraging the collected annotations for group affect, we explored two modeling paradigms for predicting group affect: (i) a handcrafted approach based on synchrony and convergence features, and (ii) a data-driven model employing graph attention mechanisms informed by social network theory. Our findings underscore the critical role of interpersonal dynamics in modeling group affect: the graph-based model consistently outperformed the handcrafted method. Moreover, integrating both audio and visual modalities improved prediction performance for arousal and valence. To further investigate group affect dynamics, we analyzed patterns of convergence and divergence using both feature-based measures and attention weights from the graph model. Quantitative results indicate that when social signals among group

members *diverged*, the group affect tended toward *neutral levels*, whereas *convergence* corresponded with more *extreme affect* (either strongly positive or negative). Our results also reveal that group members are more synchronized during the emergence of positive affect compared to negative affect.

In summary, our work advances the study of collective affect by providing a theoretically grounded annotation methodology and demonstrating that modeling fine-grained group dynamics is essential to understanding how group affect emerges and fluctuates in social interactions.

6.1 Limitations and Future Research Avenues

The group interactions present in MeMo [47] occur among groups of unacquainted participants, with a short longitudinal study spanning 3 interactions over the course of 2 weeks. While this zero-history group setup is very suitable to study emergence processes such as group affect [79], we cannot draw conclusions regarding collective affect and its convergence or divergence mechanisms in *groups that share a history*. Despite our observations of rather vivid discussions among participants in MeMo [47], the interactions are still not "real" groups that collaborate on a day-to-day basis, which may explain the relatively small nuances of affective variance in our annotations (see Fig. 1). Hence, it would be of interest to collect group affect annotations on longitudinal data of "real" groups that collaborate on a day-to-day basis as a future research endeavor.

Moreover, in line with the affect recognition literature that prioritizes *expressed* over experienced emotion [14], [17], we instructed annotators to focus solely on the affect collectively expressed by group members. However, ambiguity can arise when the experienced emotion is not fully expressed [15], such as when individuals surface-act to maintain a happy face [80], down-regulate their emotional expressions in line with social norms [81], or manage their emotional expressions for strategic purposes [82]. To capture such possibilities is beyond the scope of this research work, however this might be addressed in future research by adding self-report measures of emotional labor (e.g., [83]) following a group interaction and examine to what extent observable expressions of collective group affect as studied in the current work may be influenced by individual emotion regulation.

Finally, future work could advance large language models (LLMs) beyond individual-level affect recognition toward modeling group-level emotions and their temporal dynamics—two critical yet underexplored areas in current research. Most existing LLM-based methods remain focused on static emotion classification and lack the ability to capture affective fluctuations over time [84]–[87]. A promising direction lies in combining the relational and temporal strengths of the proposed GNN with the generative and multimodal reasoning capabilities of LLMs to better model the complex, evolving nature of group affect in social interactions. As an additional next step for LLM applications in the context of group mood, current developments also point to the opportunity to insert intelligent artificial agents into group interactions, which may help steer the group toward shared positive affect [88]. Of note however, such an agent would also need to be able to identify when divergence rather than convergence in collective group affect would be more adaptive (e.g., in order to prevent premature consensus on less than ideal solutions, cf. [89]).

REFERENCES

[1] A. Knight and N. Eisenkraft, "Positive is usually good, negative is not always bad: The effects of group affect on social integration and task performance," *Journal of Applied Psychology*, vol. 100, pp. 1214–1227, Dec. 2014.

[2] S. G. Barsade and A. P. Knight, "Group affect," *Annu. Rev. Organ. Psychol. Organ. Behav.*, vol. 2, no. 1, pp. 21–46, 2015.

[3] V. B. Hinsz and L. Bui, "Socially shared affect: Shared affect, affect sharing, and affective processing in groups." *Group Dynamics: Theory, Research, and Practice*, vol. 27, no. 4, p. 229, 2023.

[4] J. R. Hackman and N. Katz, "Group behavior and performance," *Handbook of social psychology*, vol. 2, pp. 1208–1251, 2010.

[5] S. W. Kozlowski and D. R. Ilgen, "Enhancing the effectiveness of work groups and teams," *Psychological science in the public interest*, vol. 7, no. 3, pp. 77–124, 2006.

[6] A. L. Collins, S. A. Lawrence, A. C. Troth, and P. J. Jordan, "Group affective tone: A review and future research directions," *Journal of Organizational Behavior*, vol. 34, no. S1, pp. S43–S62, 2013.

[7] C. Jones, S. Volet, and D. Pino-Pasternak, "Observational research in face-to-face small groupwork: Capturing affect as socio-dynamic interpersonal phenomena," *Small Group Research*, vol. 52, no. 3, pp. 341–376, 2021.

[8] E. A. Veltmeijer, C. Gerritsen, and K. V. Hindriks, "Automatic emotion recognition for groups: a review," *IEEE Tran. on Affective Comp.*, vol. 14, no. 1, pp. 89–107, 2021.

[9] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe, "The more the merrier: Analysing the affect of a group of people in images," in *Int. Conf. on Automatic Face and Gesture Recognition*. IEEE, 2015.

[10] Z. Lei and N. Lehmann-Willenbrock, "Affect in meetings: An interpersonal construct in dynamic interaction processes," in *The Cambridge handbook of meeting science*, 2015, pp. 456–482.

[11] G. Sharma, A. Dhall, and J. Cai, "Audio-visual automatic group affect analysis," *IEEE Tran. on Affective Comp.*, vol. 14, no. 2, 2021.

[12] J. Rösemeier, X. Hu, and B. A. Nijstad, "Toward a dynamic social process view: An integrative, multidisciplinary review of the relationship between affect and creativity," *Academy of Management Annals*, 2025.

[13] B. Ya-Hui Lien, Y.-C. Hsu, Y.-h. Chen, and L.-W. Chen, "The formation of positive group affective tone: A narrative practice," *Small Group Research*, vol. 54, no. 2, pp. 277–301, 2023.

[14] B. W. Schuller, "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.

[15] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The ambiguous world of emotion representation," *arXiv preprint arXiv:1909.00360*, 2019.

[16] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, Shanghai, China, Apr. 2013.

[17] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Tran. on Affective Comp.*, vol. 10, no. 4, pp. 471–483, Dec. 2019.

[18] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Interspeech*, Shanghai, China, Oct. 2020.

[19] X. Wang, D. Zhang, and D.-J. Lee, "Implementing the affective mechanism for group emotion recognition with a new graph convolutional network architecture," *IEEE Tran. on Affective Comp.*, 2023.

[20] X. Huang, J. Xu, W. Zheng, Q. Mao, and A. Dhall, "A survey of deep learning for group-level emotion recognition," *CoRR*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2408.15276

[21] S. G. Barsade, C. G. Coutifaris, and J. Pillemer, "Emotional contagion in organizational life," *Research in Organizational Behavior*, vol. 38, pp. 137–151, 2018.

[22] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *ACM Int. Conf. on Multimodal Interaction*, 2017.

[23] A. Dhall, M. Singh, R. Goecke, T. Gedeon, D. Zeng, Y. Wang, and K. Ikeda, "Emotiw 2023: Emotion recognition in the wild challenge," in *Proceedings of the 25th International Conference on Multimodal Interaction*, 2023, pp. 746–749.

[24] N. Lehmann-Willenbrock, "Dynamic interpersonal processes at work: Taking social interactions seriously," *Annual Review of Organizational Psychology and Organizational Behavior*, 2024.

[25] X. Huang, A. Dhall, R. Goecke, M. Pietikäinen, and G. Zhao, "Multimodal framework for analyzing the affect of a group of people," *Trans. on Multimedia*, vol. 20, no. 10, pp. 2706–2721, 2018.

[26] C. Raman, N. Raj Prabhu, and H. Hung, "Perceived conversation quality in spontaneous interactions," *IEEE Tran. on Affective Comp.*, vol. 14, no. 4, pp. 2901–2912, 2023.

[27] C. A. Bartel and R. Saavedra, "The collective construction of work group moods," *Administrative Science Quarterly*, vol. 45, no. 2, 2000.

[28] S. Hareli and A. Rafaeli, "Emotion cycles: On the social influence of emotion in organizations," *Research in organizational behavior*, vol. 28, pp. 35–59, 2008.

[29] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE/CVF Winter Conf. on App. of Computer Vision*. IEEE, 2016.

[30] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *IEEE Tran. on Affective Comp.*, vol. 3, no. 3, pp. 349–365, 2012.

[31] M. C. Nanninga, Y. Zhang, N. Lehmann-Willenbrock, Z. Szlávik, and H. Hung, "Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry," in *ACM Int. Conf. on Multimodal Interaction*, 2017, pp. 206–215.

[32] S. S. Singh, S. Muhuri, S. Mishra, D. Srivastava, H. K. Shakya, and N. Kumar, "Social network analysis: A survey on process, tools, and application," *ACM Comp. Surveys*, vol. 56, no. 8, 2024.

[33] E. R. Smith, C. R. Seger, and D. M. Mackie, "Can emotions be truly group level? evidence regarding four conceptual criteria." *Journal of personality and social psychology*, vol. 93, no. 3, p. 431, 2007.

[34] P. Ekman, *Are there basic emotions?* American Psychological Association, 1992.

[35] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[36] N. Lehmann-Willenbrock, R. A. Meyers, S. Kauffeld, A. Neininger, and A. Henschel, "Verbal interaction sequences and group mood: Exploring the role of team planning communication," *Small Group Research*, vol. 42, no. 6, pp. 639–668, 2011.

[37] W. Mou, H. Gunes, and I. Patras, "Alone versus in-a-group: A multi-modal framework for automatic affect recognition," *ACM Tran. on Multimedia Comp., Comm., and Appl.*, vol. 15, no. 2, 2019.

[38] S. G. Barsade, "The ripple effect: Emotional contagion and its influence on group behavior," *Administrative Science Quarterly*, vol. 47, no. 4, pp. 644–675, 2002.

[39] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "Emotiw 2018: Audio-video, student engagement and group-level affect prediction," in *ACM Int. Conf. on Multimodal Interaction*, 2018.

[40] S. Ghosh, A. Dhall, and N. Sebe, "Automatic group affect analysis in images via visual attribute and feature networks," in *2018 25th IEEE Int. Conf. on Image Proc.*, 2018, pp. 1967–1971.

[41] G. Sharma, S. Ghosh, and A. Dhall, "Automatic group level affect and cohesion prediction in videos," in *Int. Conf. on Affective Comp. and Intelligent Interaction Workshops and Demos*, 2019, pp. 161–167.

[42] S. Ghosh, Z. Cai, P. Gupta, G. Sharma, A. Dhall, M. Hayat, and T. Gedeon, "Emolysis: A multimodal open-source group emotion analysis and visualization toolkit," in *Int. Conf. on Affective Comp. and Intelligent Interaction Workshops and Demos*, 2024, pp. 116–118.

[43] Y. Wang, S. Zhou, Y. Liu, K. Wang, F. Fang, and H. Qian, "Congnn: Context-consistent cross-graph neural network for group emotion recognition in the wild," *Information Sciences*, 2022.

[44] X. Wang, T. Chen, and D. Zhang, "A spatial-temporal graph convolutional network for video-based group emotion recognition," in *Int. Conf. on Pattern Recognition*. Springer, 2024, pp. 339–354.

[45] E. Hatfield, J. Cacioppo, and R. Rapson, *Emotional Contagion Cambridge*. Cambridge University Press, 1994.

[46] F. Walter and H. Bruch, "The positive group affect spiral: A dynamic model of the emergence of positive affective similarity in work groups," *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, vol. 29, no. 2, pp. 239–261, 2008.

[47] M. Tsfasman, B. Dudzik, K. Fenech, A. Lorincz, C. M. Jonker, and C. Oertel, "Introducing memo: A multimodal dataset for memory modelling in multiparty conversations," *Under Review*, 2024. [Online]. Available: https://arxiv.org/abs/2409.13715

[48] N. Raj Prabhu, C. Raman, and H. Hung, "Defining and Quantifying Conversation Quality in Spontaneous Interactions," in *Comp. Pub. of ACM Int. Conf. on Multimodal Interaction*, Sep. 2020.

[49] P. Mangold, "Discover the invisible through tool-supported scientific observation," in *Böttger, H., Jensen, K., Jensen T.-Mindful Evolution. Conference Proceedings. Bad Heilbrunn: Klinkhardt.*, 2018.

[50] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, pp. 49–59, 1994.

[51] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *IEEE Int. Conf. on Affective Comp. and Intelligent Interaction*, 2017, pp. 248–255.

[52] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[53] N. Raj Prabhu, N. Lehmann-Willenbrock, and T. Gerkmann, "End-to-end label uncertainty modeling in speech emotion recognition using bayesian neural networks and label distribution learning," *IEEE Tran. on Affective Comp.*, vol. 15, no. 2, pp. 579–592, 2024.

[54] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[55] J. A. Gliem, R. R. Gliem *et al.*, "Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales," in *Midwest research-to-practice conference in adult, continuing, and community education*, vol. 1. Columbus, OH, 2003, pp. 82–87.

[56] D. Freedman, R. Pisani, and R. Purves, "Statistics," *4th Edition WW Norton & Company, New York*, 2007.

[57] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Jan. 2005, pp. 381–385.

[58] Z. Lei and N. Lehmann-Willenbrock, "Contagious peers in teams: Peer affective influence on individual emotions and performance," *Academy of Management Proceedings*, vol. 2014, Oct. 2014.

[59] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Information Fusion*, vol. 68, pp. 46–53, 2021.

[60] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, 2023.

[61] D. de Oliveira, N. Raj Prabhu, and T. Gerkmann, "Leveraging Semantic Information for Efficient Self-Supervised Emotion Recognition with Audio-Textual Distilled Models," in *Interspeech*, 2023.

[62] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," *IEEE Transactions on Affective Computing*, pp. 1–17, Jun. 2022.

[63] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Tran. on Affective Comp.*, vol. 7, no. 2, pp. 190–202, Jul. 2015.

[64] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, 2018.

[65] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[66] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, 2017, pp. 131–135.

[67] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[69] K. J. Klein and S. W. Kozlowski, "A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes," *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*, pp. 3–90, 2000.

[70] J. Vargas-Quiros, Ö. Kapcak, H. Hung, and L. Cabrera-Quiros, "Individual and joint body movement assessed by wearable sensing as a predictor of attraction in speed dates," *IEEE Tran. on Affective Comp.*, vol. 14, no. 3, pp. 2168–2181, 2023.

[71] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Int. Conf. on Learning Representations*, 2017.
[72] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Int. Conf. on Learning Representations*, 2018.
[73] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Calgary, Apr. 2018.
[74] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.
[75] J. R. Kelly and S. G. Barsade, "Mood and emotions in small groups and work teams," *Organizational behavior and human decision processes*, vol. 86, no. 1, pp. 99–130, 2001.
[76] C. Oertel, K. A. Funes Mora, J. Gustafson, and J.-M. Odobez, "Deciphering the silent participant: On the use of audio-visual cues for the classification of listener categories in group discussions," in *ACM Int. Conf. on Multimodal Interaction*, 2015, p. 107–114.
[77] S. Alisamir and F. Ringeval, "On the evolution of speech representations for affective computing: A brief history and critical overview," *IEEE Signal Proc., Magazine*, vol. 38, pp. 12–21, 2021.
[78] S. Barsade and D. Gibson, "Group affect," *Current Directions in Psychological Science*, vol. 21, pp. 119–123, 03 2012.
[79] S. W. Kozlowski, "Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations," *Organizational Psychology Review*, vol. 5, no. 4, pp. 270–299, 2015.
[80] A. S. Gabriel, M. A. Daniels, J. M. Diefendorff, and G. J. Greguras, "Emotional labor actors: A latent profile analysis of emotional labor strategies," *Journal of Applied Psychology*, vol. 100, no. 3, 2015.
[81] D. Geddes and D. Lindebaum, "Unpacking the 'why' behind strategic emotion expression at work: A narrative review and proposed taxonomy," *European Management Journal*, 2020.
[82] F. Liu and S. Maitlis, "Emotional dynamics and strategizing processes: A study of strategic conversations in top team meetings," *Journal of management studies*, vol. 51, no. 2, pp. 202–234, 2014.
[83] T. M. Glomb and M. J. Tews, "Emotional labor: A conceptualization and scale development," *Journal of Vocational Behavior*, vol. 64, no. 1, pp. 1–23, 2004.
[84] Z. Cheng, Z.-Q. Cheng, J.-Y. He, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. Hauptmann, "Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning," *Advances in Neural Inf. Proc. Sys., NeurIPS*, vol. 37, pp. 110 805–110 853, 2024.
[85] S. Dutta and S. Ganapathy, "LLM supervised pre-training for multimodal emotion recognition in conversations," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, 2025.
[86] S. Madan, S. Ghosh, L. R. Sookha, M. Ganaie, R. Subramanian, A. Dhall, and T. Gedeon, "MIP-GAF: A mllm-annotated benchmark for important person localization and group context understanding," in *IEEE/CVF Winter Conf. on App. of Computer Vision*, 2025.
[87] J. Zhang, Z. Mai, Z. Xu, and Z. Xiao, "Is LLAMA 3 good at identifying emotion? a comprehensive study," in *Proc. of Int. Conf. on Machine Learning and Machine Intelligence*, 2024, pp. 128–132.
[88] F. Koch, J. Nahulan, J. Fox, and M. Keen, "Generative intelligence systems in the flow of group emotions," *arXiv preprint arXiv:2507.11831*, 2025.
[89] M. L. To, "Emotions in action: How affective convergence and divergence contribute to team creativity," in *Organizational Behavior: Current Science, Models, and App.* Springer, 2025, pp. 701–721.

Navin Raj Prabhu received a B.Tech degree in Computer Science from SRM University, India, in 2015, and the MS degree in Computer Science from Delft University of Technology, The Netherlands, in 2020. Currently, he is a PhD student at the Signal Processing Lab and the Organization Psychology Lab, University of Hamburg, Germany. His research interests include affective computing, social signal processing, deep learning, uncertainty modeling, generative modeling, emotional speech synthesis, and group affect.

Maria Tsfasman holds a BSc degree in Fundamental and Computational linguistics from HSE university, Moscow and MSc degree in Artificial Intelligence from the Radboud University, Nijmegen with distinction. Her research interests lie between cognitive and computer science: training machines to understand humans better and using the insights from these machines to expand global understanding of human social processing and cognition.

Catharine Oertel is an Assistant Professor at TU Delft, The Netherlands. She is co-principal investigator of the Designing Intelligence Lab (DI Lab), an effort aiming to bridge research done in computer science with industrial design engineering. Her research interest lies on understanding and modeling human interaction to build socially aware conversational agents able to engage with people in a human-like manner.

Timo Gerkmann (S'08–M'10–SM'15) is a Professor for Signal Processing at the University of Hamburg, Germany. He has previously held positions at Technicolor Research & Innovation in Germany, the University of Oldenburg in Germany, KTH Royal Institute of Technology in Sweden, Ruhr-Universität Bochum in Germany, and Siemens Corporate Research in Princeton, NJ, USA. His main research interests are on statistical signal processing and machine learning for speech and audio applied to communication devices, hearing instruments, audio-visual media, and human-machine interfaces. Timo Gerkmann served as a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing (2018-2023), as an Associate Editor (2019-2022) and since 2022 serves as a Senior Area Editor of the IEEE/ACM Transactions on Audio, Speech and Language Processing. He received the VDE ITG award 2022.

Nale Lehmann-Willenbrock is Professor of Industrial and Organizational Psychology, Director of the Center for Better Work, and currently also serves as Vice Dean of Research and Transfer at the Faculty of Psychology and Human Movement Science, University of Hamburg. Previously, she held positions as associate professor at the University of Amsterdam and assistant professor at Vrije Universiteit Amsterdam. She holds a PhD in Psychology from Technische Universität Braunschweig (2012). She studies dynamic social interaction patterns in groups and teams, interpersonal processes among leaders and followers, and meetings as a core interaction site in organizations. Her research program blends organizational psychology, management, communication, and social signal processing. She served as Associate Editor for Small Group Research (2019-2024) and currently as Associate Editor for the Journal of Business and Psychology as well as Group & Organization Management.

Dynamics of Collective Group Affect: Group-level Annotations and the Multimodal Modeling of Convergence and Divergence
Appendix

Navin Raj Prabhu[†], Maria Tsfasman[⋆], Catharine Oertel[⋆], Timo Gerkmann[†], and Nale Lehmann-Willenbrock[†]

[†]University of Hamburg, Germany, [⋆]Technical University of Delft, The Netherlands.

1 Multilevel hand-crafted features extracted

Individual level	**Dyad** level		**Group** level
	Synchrony	Convergence	
Audio [1, 2] 1) Pitch 2) Speech rate 3) Intensity 4) MFCC 5) VGGish	**Origin**: [3, 4] (i) correlation coefficient (ii) lagged correlation (iii) best lag	**Origin**: [3, 5] (i) global (ii) symmetric (iii) asymmetric	**Origin**: [1, 3] (i) mean (ii) deviation (iii) median (iv) min (v) max (vi) gradient
Video [6, 7] 1) Action units 2) Facepose 3) ResNet50			

Table S1: List of the extracted hand-crafted features at individual-level, dyad-level, and group-level.

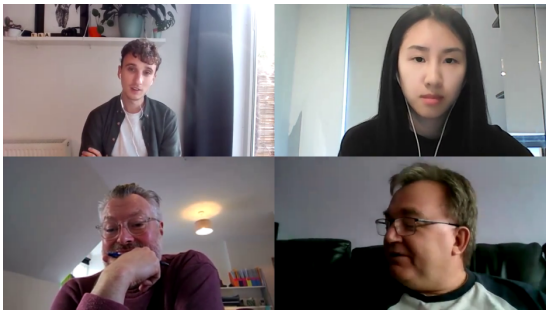
2 Snippets from the dataset

To illustrate the ebb and flow of dynamic group affect as captured by the collected annotations, Figure S1 presents video frames from time-consecutive 15-second segments, sourced from Group 12, Session 3 of the MEMO dataset [8].

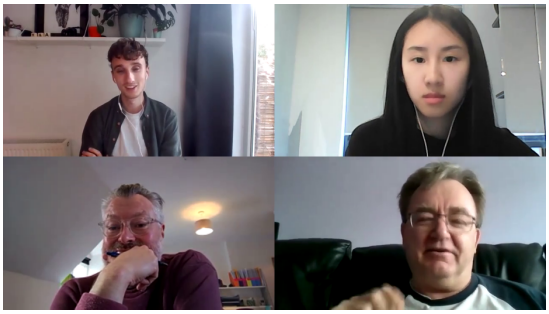
This example highlights three key strengths of the annotations in capturing both top-down and bottom-up processes of group affect. First, the overall progression of affect is clearly observed: from neutral affect in Figures S1a and S1b, rising to slightly positive in Figure S1c, peaking in Figure S1d, and then gradually returning to neutral (Figures S1e, S1f) before dipping into negative affect in Figures S1g and S1h. This smooth transition supports the idea that group affect typically flows through neutral states before shifting to more extreme levels.

Second, subtle yet meaningful differences in social signals are evident between consecutive segments. For instance, compare Figures S1c and S1d, or Figures S1g and S1h. These differences validate the precision of the annotations and demonstrate that the chosen 15-second window size is effective in capturing such nuances.

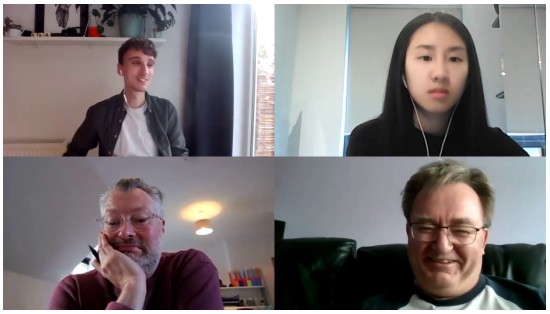
Third, the bottom-up influences are also well reflected. Comparing Figure S1a and S1b, the increase in affect is not only due to stronger individual expressions but also due to the number of group members displaying the cues. In Figure S1a, only one person shows a faint smile, corresponding to a neutral affect annotation. In contrast, in Figure S1b, two members (half the group) smile, resulting in a slight rise in arousal and valence. This pattern also holds for negative affect—see Figures S1g and S1h.



(a) Onset time: 09:30 (min:sec)
Ground-truth: Arousal 0.52 and Valence 0.60
Predicted: Arousal 0.58 and Valence 0.59



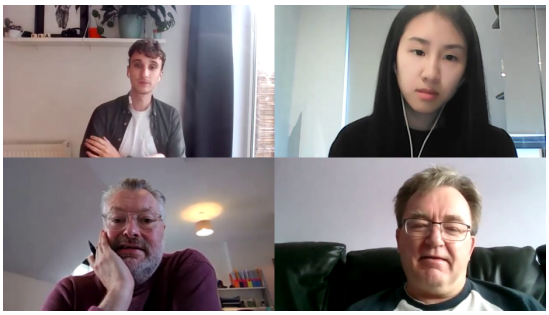
(b) Onset time: 09:45 (min:sec)
Ground-truth: Arousal 0.60 and Valence 0.60
Predicted: Arousal 0.56 and Valence 0.59



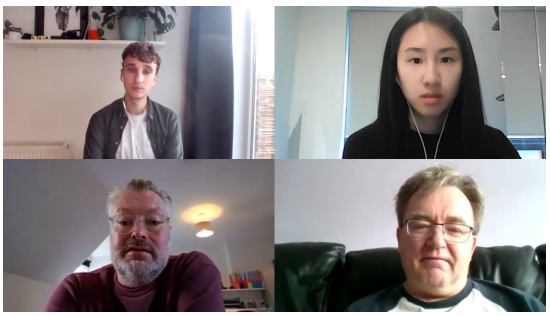
(c) Onset time: 10:00 (min:sec)
Ground-truth: Arousal 0.65 and Valence 0.65
Predicted: Arousal 0.66 and Valence 0.66

(d) Onset time: 10:15 (min:sec)
Ground-truth: Arousal 0.85 and Valence 0.85
Predicted: Arousal 0.78 and Valence 0.85

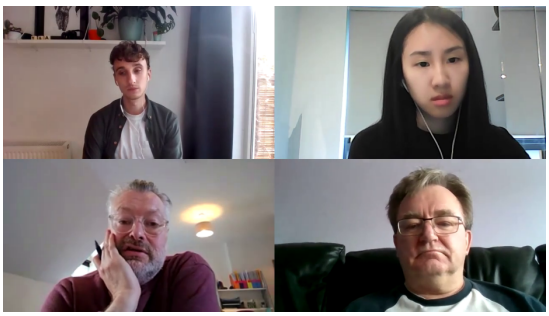
(e) Onset time: 10:30 (min:sec)
Ground-truth: Arousal 0.55 and Valence 0.53
Predicted: Arousal 0.59 and Valence 0.57



(f) Onset time: 10:45 (min:sec)
Ground-truth: Arousal 0.40 and Valence 0.47
Predicted: Arousal 0.50 and Valence 0.54



(g) Onset time: 11:00 (min:sec)
Ground-truth: Arousal 0.37 and Valence 0.37
Predicted: Arousal 0.48 and Valence 0.46



(h) Onset time: 11:15 (min:sec)
Ground-truth: Arousal 0.35 and Valence 0.32
Predicted: Arousal 0.38 and Valence 0.35

Figure S1: Sampled video frame from each consecutive 15-second segment of **Group 12 Session 3** in the MEMO dataset [8].

3 Comparison with Existing Group Affect Datasets

	Group Interaction Format & Capture Method			Annotator Information			Annotation Characteristics		Dataset Details	
	Modality	Purposive	Examples	#	Training	Background	Dynamics	Window Size	Sample Size	Ground-truth
HAPPEI (2012) [9, 10]	Image	✗	Social events: party, marriage, convocation etc.,	4	✗	Naive	✗	✗	2,638 images	Happiness Intensity (6 levels)
MultiEmoVA (2015) [11]	Image	✗	Social events: sport events, sport, crowd etc.,	**14**	✗	Graduate Students	✗	✗	250 images	Arousal and Valence (3 levels each)
DynamicAffect (2015) [12]	**Video**	✓	**Team Meetings**	2	✓ **(Video Markers based Training)**	Graduate Students	✓	2 mins	500 videos (2 mins segments)	**Circumplex Model: Arousal and Valence (continuous values)**
GAF 2.0 (2015) [13]	Image	✗	Social events: party, marriage, convocation etc.,	3	✗	Naive	✗	✗	6,467 images	3-levels of affect (positive, negative, neutral)
GAF 3.0 (2018) [14]	Image	✗	Social events: party, marriage, convocation etc.,	3	✗	Naive	✗	✗	17,172 images	3-levels of Valence
Group Cohesion (2019) [15]	Image	✗	Social events: wedding, family, laughing club, birthday party etc.,	5	✗	Naive	✗	✗	16,433 images	3-levels of Valence
AloneVsGroup (2019) [16]	**Video**	✗	Group of people engaging with multimedia content	3	✗	Naive	✓	20 secs	7,630 videos (20 secs segments)	**Circumplex Model: Arousal and Valence (continuous values)**
VGAF (2020) [17]	**Video**	✗	YouTube Videos: protest, festival, wedding, fighting, etc.,	3	✓ (Informed with concepts of emotion)	Naive	✗	✗	4,183 videos (≈5 secs each)	3-levels of Valence
GroupEmoW (2020) [18]	Image	✗	Social events: party, protest, wedding etc.,	5	✗	Students and Professors	✗	✗	15,894 images	3-levels of Valence
SiteGroEmo (2022) [19]	Image	✗	Groups in travel destination	5	✗	With prior knowledge on emotion	✗	✗	10,034 images	3-levels of Valence
GECV (2022) [20]	**Video**	✗	Crowd videos: festival, marching, wedding party, parade, funeral, etc.,	3	✗	Naive	✗	✗	627 videos (20 secs each)	3-levels of Valence
MIP-GAF (2025) [21]	Image	✗	Casual Gathering, Celebration, Funeral, Concerts, Wrestling, etc..,	2 + 1 LLM	✗	Naive	✗	✗	16,550	Most Important Person
Ours	**Video**	✓	**Team Meetings** [8]	**8**	✓ **(Video Markers based Training)**	**Psychology Graduate Students**	✓	**15 secs (Iteratively Tuned)**	3,794 videos (15 secs each)	**Circumplex Model: Arousal and Valence (continuous values)**

Table S2: Comparison of introduced annotations with existing annotation schemes in the literature

4 Overall experimental workflow of research

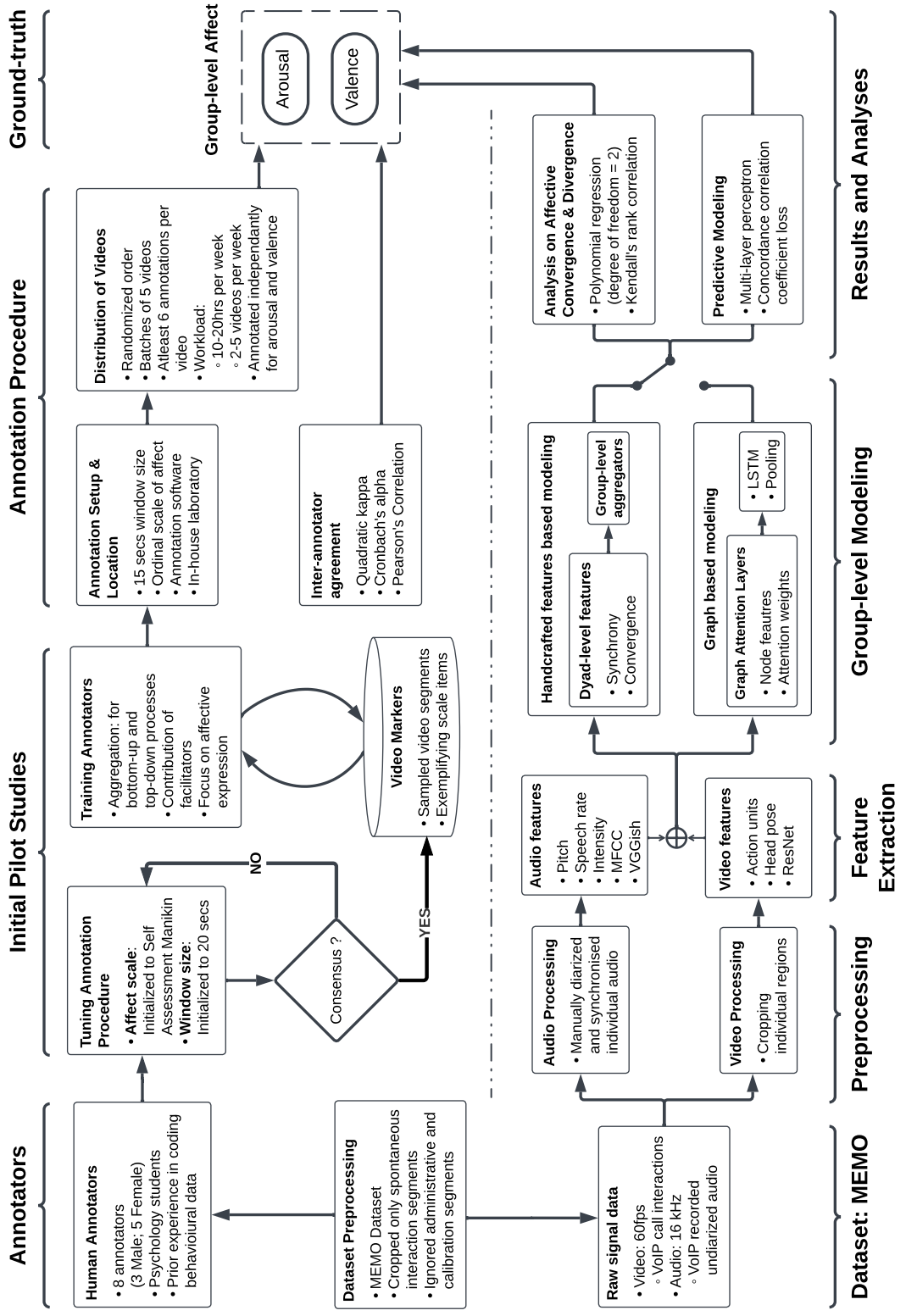
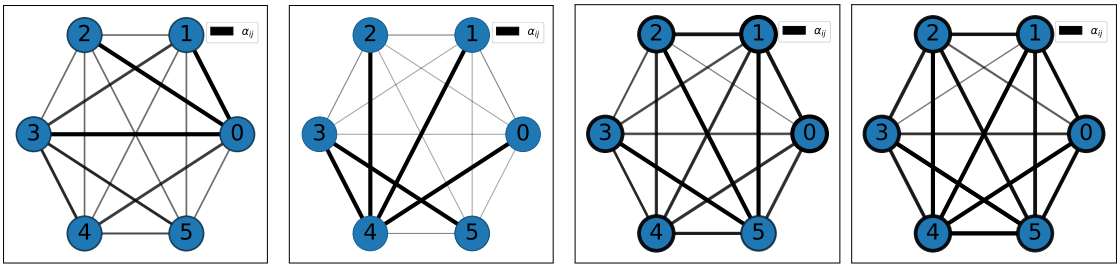


Figure S2: Overview of the experimental workflow followed in this research work.

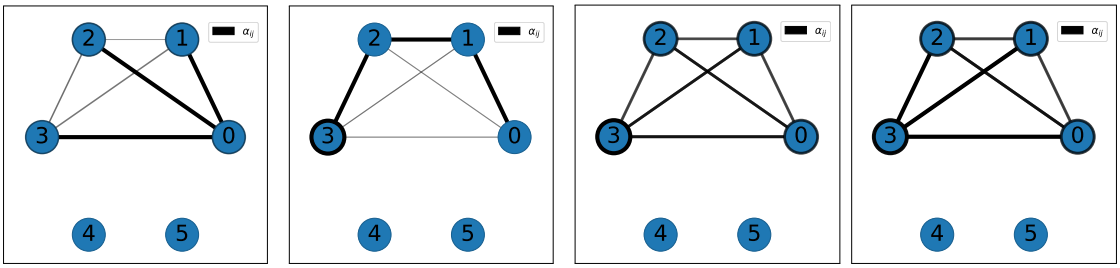
5 Supplementary analyses: Graph based modeling

5.1 Graph visualization for Positive, Negative and Neutral affect



(a) Arousal:0.66, Valence:0.70 **(b)** Arousal:0.21, Valence:0.42 **(c)** and (d): Arousal:0.55, Valence:0.52

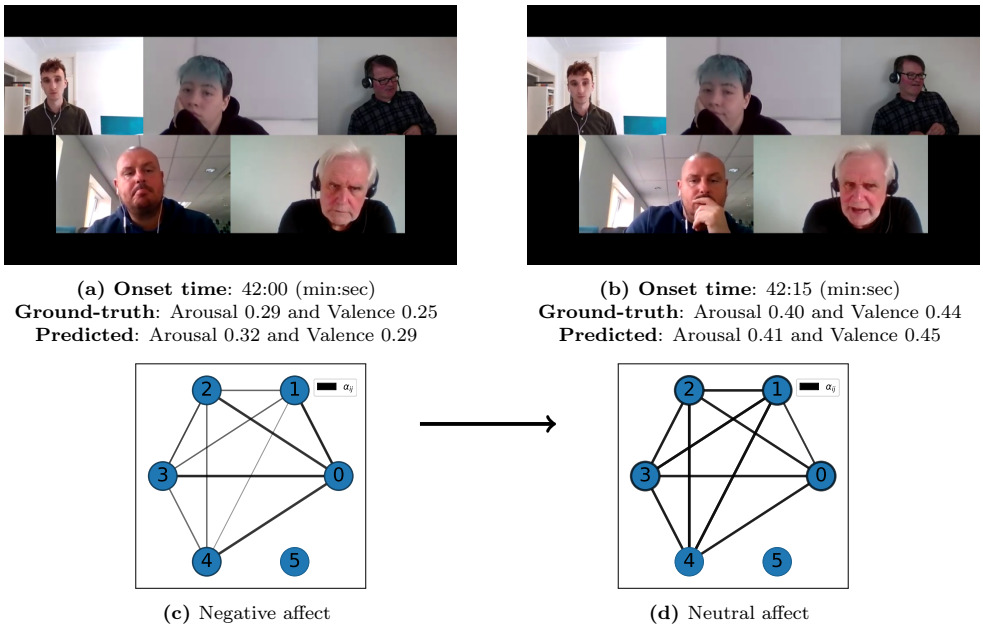
Figure S3: Visualization of attention weights α_{ij}, for Positive (a), Negative (b), and Neutral (c and d) samples. Sampled from the **Group 10 Session 2** interaction of the MEMO dataset [8].



(a) Arousal:0.87, Valence:0.70 **(b)** Arousal:0.43, Valence:0.27 **(c)** and (d): Arousal:0.50, Valence:0.50

Figure S4: Visualization of attention weights α_{ij}, for Positive (a), Negative (b), and Neutral (c and d) samples. Sampled from the **Group 12 Session 1** interaction of the MEMO dataset [8].

5.2 Graph evolution across consecutive time segments

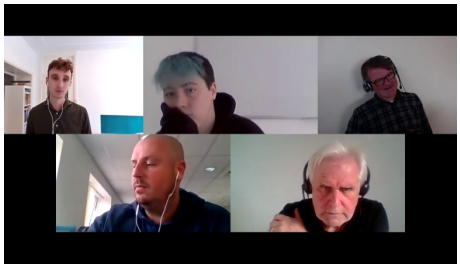


(a) Onset time: 42:00 (min:sec)
Ground-truth: Arousal 0.29 and Valence 0.25
Predicted: Arousal 0.32 and Valence 0.29

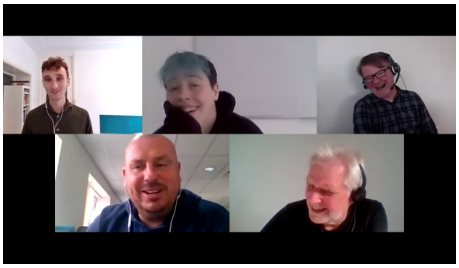
(b) Onset time: 42:15 (min:sec)
Ground-truth: Arousal 0.40 and Valence 0.44
Predicted: Arousal 0.41 and Valence 0.45

(c) Negative affect **(d)** Neutral affect

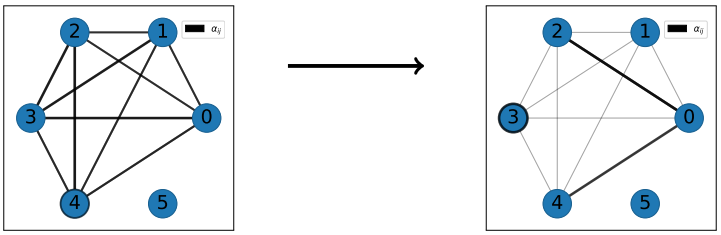
Figure S5: Evolution of the graph structure from **Negative** to **Neutral** group affect in the bottom row (c, d), and corresponding 15-second segments in the top row (a, b).



(a) Onset time: 45:00 (min:sec)
Ground-truth: Arousal 0.65 and Valence 0.65
Predicted: Arousal 0.66 and Valence 0.66



(b) Onset time: 45:15 (min:sec)
Ground-truth: Arousal 0.85 and Valence 0.85
Predicted: Arousal 0.78 and Valence 0.85



(c) Moderately Positive affect

(d) Extreme Positive affect

Figure S6: Evolution of the graph structure from **Moderately Positive** to **Extreme Positive** group affect in the bottom row (c, d), and corresponding 15-second segments in the top row (a, b).

Figures S5 and S6 illustrate moments where transitions between different affect levels occur across consecutive segments, which subsequently trigger changes in graph attention weights. For instance, Figure S5 illustrates a transition from Negative to Neutral affect. Based on the two visualized transitions, we draw the following inferences.

First, when affect transitions from *Negative* to *Neutral* levels (see FigureS5c and Figure S5d), the graph evolves from exhibiting high deviation in attention weights $\sigma(\alpha_{ij})$ and low self-attention α_{ii} to lower $\sigma(\alpha_{ij})$ and higher α_{ii}. Second, when affect levels transition from moderate to extreme, for instance from *Moderately Positive* to *Strongly Positive* (see FigureS6c and Figure S6d), $\sigma(\alpha_{ij})$ also increases. This indicates a potential linear correlation between affect intensity and the corresponding $\sigma(\alpha_{ij})$. Notably, a distinction still exists between Neutral (see Figure S5d) and Moderately Positive (see Figure S6c) levels, where self-attention α_{ii} values are higher for Neutral affect than for Moderately Positive affect.

Overall, these findings suggest that at extreme levels of affect, modeling is more effectively driven by interpersonal relationships, indicating greater group synchrony—whether convergent or divergent—compared to neutral states. In contrast, during neutral affect, group members appear less synchronous, with individual-level cues within the segment playing a more prominent role in characterizing the affect.

References

[1] M. C. Nanninga, Y. Zhang, N. Lehmann-Willenbrock, Z. Szlávik, and H. Hung, "Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry," in *ACM Int. Conf. on Multimodal Interaction*, 2017, pp. 206–215.

[2] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Tran. on Affective Comp.*, vol. 7, no. 2, pp. 190–202, Jul. 2015.

[3] C. Raman, N. Raj Prabhu, and H. Hung, "Perceived conversation quality in spontaneous interactions," *IEEE Tran. on Affective Comp.*, vol. 14, no. 4, pp. 2901–2912, 2023.

[4] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *IEEE Tran. on Affective Comp.*, vol. 3, no. 3, pp. 349–365, 2012.

[5] J. Vargas-Quiros, Ö. Kapcak, H. Hung, and L. Cabrera-Quiros, "Individual and joint body movement assessed by wearable sensing as a predictor of attraction in speed dates," *IEEE Tran. on Affective Comp.*, vol. 14, no. 3, pp. 2168–2181, 2023.

[6] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE/CVF Winter Conf. on App. of Computer Vision*. IEEE, 2016.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] M. Tsfasman, B. Dudzik, K. Fenech, A. Lorincz, C. M. Jonker, and C. Oertel, "Introducing memo: A multimodal dataset for memory modelling in multiparty conversations," *Under Review*, 2024. [Online]. Available: https://arxiv.org/abs/2409.13715

[9] A. Dhall, J. Joshi, I. Radwan, and R. Goecke, "Finding happiest moments in a social context," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 613–626.

[10] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "Emotiw 2016: video and group-level emotion recognition challenges," in *ACM Int. Conf. on Multimodal Interaction*. New York, NY, USA: Association for Computing Machinery, 2016, p. 427–432. [Online]. Available: https://doi.org/10.1145/2993148.2997638

[11] W. Mou, O. Celiktutan, and H. Gunes, "Group-level arousal and valence recognition in static images: Face, body and context," in *11th International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 5. IEEE, 2015, pp. 1–6.

[12] Z. Lei and N. Lehmann-Willenbrock, "Affect in meetings: An interpersonal construct in dynamic interaction processes," in *The Cambridge handbook of meeting science*, 2015, pp. 456–482.

[13] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe, "The more the merrier: Analysing the affect of a group of people in images," in *Int. Conf. on Automatic Face and Gesture Recognition*. IEEE, 2015.

[14] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "Emotiw 2018: Audio-video, student engagement and group-level affect prediction," in *ACM Int. Conf. on Multimodal Interaction*, 2018.

[15] S. Ghosh, A. Dhall, N. Sebe, and T. Gedeon, "Predicting group cohesiveness in images," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[16] W. Mou, H. Gunes, and I. Patras, "Alone versus in-a-group: A multi-modal framework for automatic affect recognition," *ACM Tran. on Multimedia Comp., Comm., and Appl.*, vol. 15, no. 2, 2019.

[17] A. Dhall, G. Sharma, R. Goecke, and T. Gedeon, "Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 784–789.

[18] X. Guo, L. Polania, B. Zhu, C. Boncelet, and K. Barner, "Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases," in *IEEE/CVF Winter Conf. on App. of Computer Vision*, 2020, pp. 2921–2930.

[19] Y. Wang, S. Zhou, Y. Liu, K. Wang, F. Fang, and H. Qian, "Congnn: Context-consistent cross-graph neural network for group emotion recognition in the wild," *Information Sciences*, 2022.

[20] K. G. Quach, N. Le, C. N. Duong, I. Jalata, K. Roy, and K. Luu, "Non-volume preserving-based fusion to group-level emotion recognition on crowd videos," *Pattern Recognition*, vol. 128, p. 108646, 2022.

[21] S. Madan, S. Ghosh, L. R. Sookha, M. Ganaie, R. Subramanian, A. Dhall, and T. Gedeon, "MIP-GAF: A mllm-annotated benchmark for important person localization and group context understanding," in *IEEE/CVF Winter Conf. on App. of Computer Vision*, 2025.

4

Synthesis of Emotional Expressions

4.1 Speech Emotion Conversion using Neural Vocoder [P9]

Abstract

Speech emotion conversion aims to convert the expressed emotion of a spoken utterance to a target emotion while preserving the lexical information and the speaker's identity. In this work, we specifically focus on in-the-wild emotion conversion where parallel data does not exist, and the problem of disentangling lexical, speaker, and emotion information arises. In this paper, we introduce a methodology that uses self-supervised networks to disentangle the lexical, speaker, and emotional content of the utterance, and subsequently uses a HiFiGAN vocoder to resynthesise the disentangled representations to a speech signal of the targeted emotion. For better representation and to achieve emotion intensity control, we specifically focus on the arousal dimension of continuous representations, as opposed to performing emotion conversion on categorical representations. We test our methodology on the large in-the-wild MSP-Podcast dataset. Results reveal that the proposed approach is aptly conditioned on the emotional content of input speech and is capable of synthesising natural-sounding speech for a target emotion. Results further reveal that the methodology better synthesises speech for mid-scale arousal (2 to 6) than for extreme arousal (1 and 7).

Reference

N. Raj Prabhu and N. Lehmann-Willenbrock and T. Gerkmann, "In-the-wild Speech Emotion Conversion Using Disentangled Self-Supervised Representations and Neural Vocoder-based Resynthesis", *Speech Communication, 15th ITG Conference*, Aachen, Germany, September, 2024, pp. 176-180, DOI: 10.30420/456164034.

Copyright Notice

Authors' Contributions

Navin Raj Prabhu led the study, including the initial conceptualization, algorithm development, neural network training, experimental validation, and manuscript preparation. Nale Lehmann-Willenbrock contributed by reviewing the manuscript and helping to refine the argumentation and overall framing. Timo Gerkmann provided key insights into the experimental validation, offered valuable methodological feedback through discussions, and participated in the manuscript review.

In-the-wild Speech Emotion Conversion Using Disentangled Self-Supervised Representations and Neural Vocoder-based Resynthesis

Navin Raj Prabhu[⋆†], *Nale Lehmann-Willenbrock*[†], *Timo Gerkmann*[⋆]

[⋆]Signal Processing, [†]Industrial and Organizational Psychology, Universität Hamburg, Germany
Email: {navin.raj.prabhu, nale.lehmann-willenbrock, timo.gerkmann}@uni-hamburg.de

Abstract

Speech emotion conversion aims to convert the expressed emotion of a spoken utterance to a target emotion while preserving the lexical information and the speaker's identity. In this work, we specifically focus on in-the-wild emotion conversion where parallel data does not exist, and the problem of disentangling lexical, speaker, and emotion information arises. In this paper, we introduce a methodology that uses self-supervised networks to disentangle the lexical, speaker, and emotional content of the utterance, and subsequently uses a HiFiGAN vocoder to resynthesise the disentangled representations to a speech signal of the targeted emotion. For better representation and to achieve emotion intensity control, we specifically focus on the arousal dimension of continuous representations, as opposed to performing emotion conversion on categorical representations. We test our methodology on the large in-the-wild MSP-Podcast dataset. Results reveal that the proposed approach is aptly conditioned on the emotional content of input speech and is capable of synthesising natural-sounding speech for a target emotion. Results further reveal that the methodology better synthesises speech for mid-scale arousal (2 to 6) than for extreme arousal (1 and 7).

1 Introduction

One way in which emotions are expressed by individuals in social interactions is via speech signals [1]. In the context of human-machine interaction systems, the generation of spoken dialogue is a fundamental facet of natural interaction between humans and machines [2, 3]. More importantly, to improve the naturalness of machine communication, the generation of emotionally expressive speech is required. While speech generation technologies have been making significant progress [4, 5], emotionally expressive speech generation is still a challenge [6, 7]. Speech emotion conversion (SEC) is a technique that aims to convert the expressed emotion of a spoken utterance to a target emotion while preserving the lexical and the speaker information [6, 8]. Therefore, SEC has a crucial application in building next-generation human-machine interaction systems, aiming at equipping them with the ability to interact with social and emotional intelligence.

Emotion can be represented either *categorically* or as *dimensional* representations. As *categorical* representations, Ekman's six basic emotions [9] (e.g., anger, happy) are commonly used [10]. However, emotion is a fuzzy construct with *fuzzy* class boundaries [11], and such discrete representations do not aptly capture the subtle difference between human emotions [11]. To overcome this, the circumplex model [11] captures emotional expressions using two continuous and independent dimensions, i.e., *arousal* (relaxed or passive vs. aroused or activated) and *valence* (positive vs. negative) [12, 13]. In SEC research, efforts have also been made to control the intensity of categorical emotion representations, e.g., using mixed emotion representations [3] or modeling emotion intensity as an auxiliary task [14]. Note that SEC using the dimensional representations directly archives intensity control, as opposed to an additional effort in the categorical representation case.

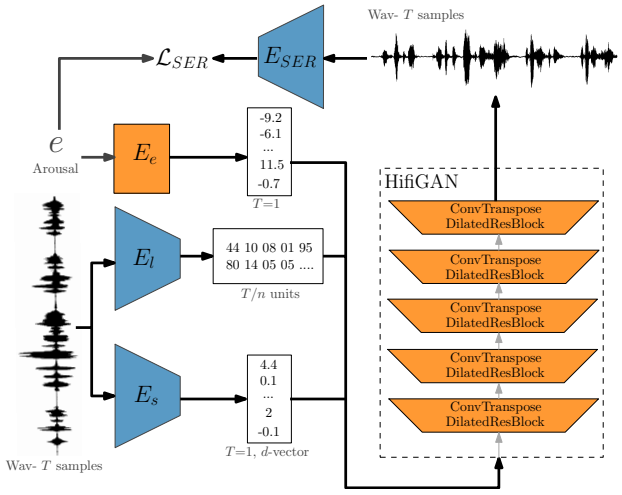


Figure 1: Illustration of the training process. During inference, e is replaced by the target emotion $\bar{e}$. *Orange* blocks denote trainable blocks, and *blue* blocks are pretrained.

SEC datasets are primarily *acted-out* [7, 10, 15, 16], as opposed to *in-the-wild* improvised datasets [12]. Acted-out databases make a strong assumption on the availability of parallel utterances [6], i.e., one where each source utterance also has a ground-truth utterance of a target emotion, which is resource inefficient to collect [17, 18] and techniques relying on them lack scalability [6].

In this work, we specifically focus on non-parallel data. Models trained on non-parallel data are scalable to different emotion types [6], as they are not restricted by the emotion pairs trained on. However, they are also more challenging than modeling parallel data, as the problem of *disentanglement* arises [6, 8]. For SEC on non-parallel data, a disentanglement method is required to decompose the input speech signal into several constituents: emotion, lexical, and speaker information, encoded in respective latent representations. During inference, the latent representation encoding emotion is modified to achieve emotion conversion. Existing works have primarily employed variational auto-encoders (VAE) [15, 19] and sequence-to-sequence encoder-decoders [3, 14] to achieve disentanglement.

Self-supervised learning (SSL) techniques that leverage large unlabeled datasets have shown to be promising in several tasks, most importantly its state-of-the-art performance in speech emotion recognition (SER) [20, 21] and speaker voice conversion [22]. Motivated by this, in this paper, we propose a novel methodology that employs SSL models to disentangle emotion, lexical, and speaker iden-

tity representations, and uses a HiFiGAN [4] to resynthesise the disentangled representations into emotional speech. For a better representation of emotion and better intensity control in emotion conversion, in this paper, we will represent emotion using the continuous arousal dimension. Moreover, we train and validate the proposed methodology on the in-the-wild MSP-Podcast v1.10 dataset [12]. To the best of our knowledge, this work is the first in the literature to perform SEC on an in-the-wild setting and on arousal, while existing literature has primarily focused on categorical emotion (e.g., anger, sad, neutral, happy) [15, 17, 18] and acted-out datasets (e.g., IEMOCAP, ESD) [7, 10].

2 Background

2.1 Disentangled SSL representations

SSL techniques have shown great success in several downstream tasks such as automatic speech recognition [23], phoneme segmentation [24], speaker verification [25], and SER [20, 21]. Existing literature well documents that representations obtained from SSL models finetuned on a specific task have exclusive information on that particular task. Analysis presented in [26] reveals that discrete representations obtained from HuBERT contain exclusively phoneme and lexical information, without any speaker or F0 information. In [25], the HuBERT framework was used to build the WavLM which was then fine-tuned exclusively to preserve speaker identity [25]. Motivated by this, in this paper, we use SSL networks to disentangle lexical and speaker information from input speech.

2.2 Speech resynthesis

While disentanglement is only the first step of SEC, the subsequent step is to synthesise the disentangled SSL representations into natural-sounding speech that is also conditioned on the emotion content. Traditionally, neural network based vocoders conditioned on the log *Mel-spectogram* enabled generating natural-sounding speech [27]. More recently, techniques have been proposed to directly synthesise natural-sounding speech from SSL-based speech representations [5, 22, 26, 28], the technique termed as *resynthesis*. In [22], HiFiGAN was used to resynthesise speech from disentangled representations obtained from three SSL models: HuBERT for lexical content, VQ-VAE for F0 information, and a speaker verification model. In this paper, we follow a similar approach, that is to perform resynthesis by training a HiFiGAN on disentangled SSL representations. The crucial difference between our work and [22] is that we employ this technique in SEC, thereby also requiring us to condition the HiFiGAN on emotion information.

3 Proposed Methodology

The task of SEC is formulated as: given a single-channel audio input of a spoken utterance $\mathbf{X}_{l,s,e} \in \mathbb{R}^{1\times T}$, with lexical content l, speaker identity s, and annotated arousal ground-truth e, with its raw signal represented as a sequence of samples $x = (x_1, ..., x_T)$, we design a system to synthesise $\widehat{\mathbf{Y}}_{l,s,\bar{e}} \in \mathbb{R}^{1\times T}$, a single-channel audio that preserves the lexical content l and speaker identity s of $\mathbf{X}_{l,s,e}$, and converts the arousal of $\mathbf{X}_{l,s,e}$ to target arousal $\bar{e}$. To this, we introduce an SSL-based HiFiGAN network (see Fig. 1) that consists of (i) *SSL encoders* that disentangles lexical l, speaker s, and emotion e information, and (ii) a *HiFiGAN decoder* that is trained to resynthesise emotion converted speech $\widehat{\mathbf{Y}}_{l,s,\bar{e}}$ from the disentangled representations. The network does not rely on parallel data, where the ground-truth $\mathbf{Y}_{l,s,\bar{e}}$ exists, and is generative in nature.

3.1 Disentanglement encoders

Lexical Encoder: The input to the lexical encoder E_l is the time-domain signal x, and its output is a sequence of low frequency representations $E_l(x) = (l_1, ..., l_{T'})$. As E_l, we choose the pretrained SSL-based HuBERT model. Following [22], using the k-means algorithm, we convert the continuous HuBERT representations into discrete representations denoted as $z_l = (z_1, ..., z_{T'})$, where each unit z_i is a positive integer and $z_i \in 0, 1, .., K$. Existing works [5, 22], have shown such discrete representations z_i to be more related to the phonemes of the respective utterance.
Speaker Encoder: Similar to E_l, the input to the speaker encoder E_s is the time-domain signal x. As E_l, the pretrained WavLM speaker verification model [25] is used. The output of E_s is a continuous $d-vector$ speaker representation $z_s \in \mathbb{R}^{512}$. Unlike the lexical representation z_l, z_s is a global representation that encodes speaker information of the whole utterance. To account for the mismatch in frequency, z_s is concatenated on all T' frames of z_l for the resynthesis and is denoted as $z_{T'} = (z_l, z_s)$.
Emotion encoder: To encode emotional information in the disentanglement phase, as the emotion encoder E_e we use simple *trainable* linear layers. The input to the encoder E_e is the ground-truth annotated arousal scalar label $e \in \mathbb{R}$ and $1 \leq e \leq 7$, and the output is the emotion representation $z_e \in \mathbb{R}^{128}$. Similar to z_s, z_e is also a global representation that encodes speaker information of the whole utterance x, and similarly z_e is also concatenated on all frames of z_l for the resynthesis and is denoted as $z_{T'} = (z_l, z_s, z_e)$.

3.2 Resynthesis decoder

Inspired by existing works on voice conversion [5, 22] and SEC [28], we adopt a modified version of the HiFiGAN [4] to resynthesise the disentangled encoder representations. Note that [28] is the closest to our work. However, in this work, we extend the HiFiGAN to *non-parallel* data and also tackle the thereby arising problem of disentanglement.

The HiFiGAN is trained using a generator G and two discriminator networks, the multi-*period* and the multi-*scale* discriminators, D_p and D_s, respectively [4, 22]. The generator G has a series of blocks composed of a transposed convolution and a dilated residual layer (see Fig. 1). The transposed convolutions upsample the representations to match the input number of samples T. The dilated residual layers increase the receptive field. The input to G is the concatenated disentangled representations $z_{T'} = (z_l, z_s, z_e)$ and the output is the resynthesised speech $\widehat{\mathbf{Y}}_{l,s,\bar{e}}$ conditioned on lexical, speaker and emotion information.

The multi-period discriminator D_p consists of six sub-discriminators each operating on different *period hops* of the input and generated speech: 2, 3, 4, 5, 7, and 11. Similarly, the multi-scale discriminator D_s uses three sub-discriminators operating at different *scales*: the original scale, x2 downsampled scale, and x4 downsampled scale.

3.3 Loss functions

HiFiGAN is a generative network trained using adversarial learning strategy [4]. Following [22], for resynthesised

output speech signal $\widehat{\mathbf{Y}}$ represented as $\hat{y} = (\hat{y}_1, ..., \hat{y}_T)$, each of the sub-discriminators D_j in D is trained to minimize the following adversarial losses (L_{adv} and L_D):

$$L_{adv}(D_j, G) = \sum_x ||1 - D_j(\hat{y})||_2^2, \quad (1)$$

$$L_D(D_j, G) = \sum_x [||1 - D_j(x)||_2^2 + ||D_j(\hat{y})||_2^2], \quad (2)$$

where x is the input speech and $\hat{y} = G(z_{T'}) = G(z_l, z_s, z_e)$.

Along with L_{adv} and L_D, a reconstruction loss term L_{recon} and a feature-matching loss L_{fm} [29] is also added to the loss function. L_{recon} measures the Mel-spectrogram reconstruction between input x and resynthesised output $\hat{y}$:

$$L_{recon}(G) = \sum_x ||\phi(x) - \phi(\hat{y})||_1, \quad (3)$$

where ϕ is a function computing Mel-spectrogram. The L_{fm} term is the distance between the discriminator activations of input x and resynthesised output $\hat{y}$:

$$L_{fm}(D_j, G) = \sum_x \sum_{i=1}^{R} \frac{1}{M_i} ||\psi_i(x) - \psi_i(\hat{y})||_1, \quad (4)$$

where ψ_i is the function that extracts the activations of the i-th discriminator layer, and, M_i and R are the number of features in i and the number of layers in D_j, respectively.

While the concatenation of emotion label embeddings z_e to the input of generator G already conditions the resynthesised output $\hat{y}$, we also included an SER loss to the loss function. To achieve this, we use a pre-trained SSL-based SER system E_{SER} introduced in [21]. The E_{SER} network was built by fine-tuning the Wav2Vec2-Large-Robust network [23] on the MSP-Podcast (v1.7) dataset [12]. The SER loss term L_{SER} is formulated as,

$$L_{SER} = \sum_x [1 - L_{ccc}(e, E_{SER}(\hat{y}))], \quad (5)$$

where L_{ccc} is the concordance correlation coefficient (CCC) [30] that measures similarity between two variables, e is the ground-truth emotion of input, and $E_{SER}(\hat{y})$ is the predicted emotion for resynthesised speech $\hat{y}$. The L_{ccc} measure varies between -1 and $+1$, where $+1$ denotes perfect similarity, therefore $1 - L_{ccc}$ is minimized during training.

The final loss for generator G and discriminator D is:

$$L_G(D, G) = \sum_{j=1}^{J} [L_{adv}(D_j, G) + \lambda_{fm} L_{fm}(D_j, G)] + \lambda_r L_{recon}(G) + \lambda_{SER} L_{SER}, \quad (6)$$

$$L_D(D, G) = \sum_{j=1}^{J} L_D(D_j, G), \quad (7)$$

where J is the number of sub-discriminators in D. Following [22], we set $\lambda_{fm} = 2$ and $\lambda_r = 45$. λ_{SER} was set to 1 after preliminary results supported this setup.

4 Experimental Setup

Dataset: The dataset used in this work is the *in-the-wild* MSP-Podcast dataset (v1.10) [12]. To the best of our knowledge, this is the first work in literature that performs SEC on an in-the-wild dataset. The dataset consists of approximately 166hrs of audio collected from podcasts. Emotion is labeled at the utterance-level in terms of arousal, valence, and dominance. In this work, we only use the arousal annotations for SEC. The arousal annotations, collected on a scale of 1 to 7, are distributed with μ =4 and σ =0.95, denoting that there are comparatively fewer samples on the extremes of the scales (1 and 7) than on middle regions (3 to 5). Note this also reflects the nature of emotions in-the-wild podcasts and also in any real-world scenario, where extreme emotions are expected to be sparse.

Validation measures: We validate the proposed methodology in terms of both the SEC capabilities and the naturalness of synthesised speech. As the measure of SEC capability, we use the mean-squared L_{mse} and mean-absolute L_{abs} errors, calculated between the target arousal $\bar{e}$ and the SER prediction on the resynthesised output $E_{SER}(\hat{y})$. As the measure of the naturalness of $\hat{y}$, we use the wav2vec mean-opinion score (WVMOS) [31]. WVMOS is an objective speech quality measure based on wav2vec2.0 [23] and is fine-tuned on the mean-opinion scores (on a scale of 1 to 5) obtained from the listening tests of the 2018 Voice Conversion Challenge [32]. As the listening tests primarily focused on the naturalness of the voice conversion task, it makes the WVMOS well suited for non-intrusively measuring the naturalness of $\hat{y}$. Note here that, as we do not use parallel data we can only rely on non-intrusive measures that do not require the ground-truth audio of emotion conversion $\mathbf{Y}_{l,s,\bar{e}}$. Statistical significance for improved performance is estimated using one-tailed t-test on error distributions, asserting significance for p-values ≤ 0.05.

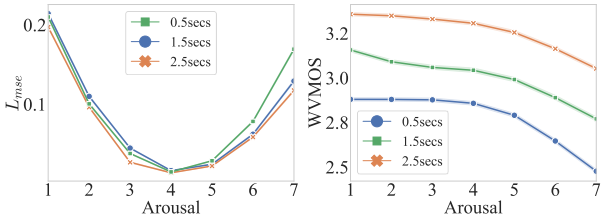
Model versions	WVMOS	SER Error	
		L_{mse}	L_{abs}
z_l	1.80	–	–
$z_l + z_s$	2.29	–	–
$z_l + z_s + z_e$	2.64	0.0971	0.2642
$z_l + z_s + z_e + L_{SER}$	**3.26***	**0.0843***	**0.2442***

Table 1: Overall performance of model versions. * indicates statistically significant improvements in results.

5 Results and Discussion

Overall performance: Firstly, we compare the overall performance of four different versions of the proposed methodology: (i) z_l, which only uses the lexical representations for resynthesis, (ii) $z_l + z_s$, which also uses the speaker information z_s, (iii) $z_l + z_s + z_e$, which also uses the emotion label embeddings z_e thereby also conditioning on emotion information, and (iv) $z_l + z_s + z_e + L_{SER}$ which further conditions on emotion information by including L_{SER} in the loss function. As this is the first work in literature to perform SEC in terms of *arousal*, we do not have any baselines, rather we present different versions of the proposed methodology for comparison.

From the results presented in Table 1, we note the following: firstly, with the increasing addition of disentangled representations the HiFiGAN is capable of producing more natural-sounding speech. The usage of all three representations along with the inclusion of SER loss $z_l + z_s + z_e + L_{SER}$ achieves the best WVMOS score of 3.26 with *statistical significance*. This also validates the SEC capability of the proposed approach, as conditioning on emotion should also improve the naturalness of synthesised speech and it is reflected in the WVMOS score. In [22], it is suggested



(a) L_{mse} trends (b) WVMOS trends

Figure 2: Class-wise performance for target arousal $\bar{e}$, with respect to the input segment sizes.

to also include F0 representations to obtain better natural-sounding speech. However, in our case, we do not use F0 representations for resynthesis as conditioning on F0 will also affect the conditioning on emotion representations z_e, given that F0 and emotion are highly correlated [33]. Secondly, including the SER loss term L_{SER} for further conditioning on emotional content results in an improved SEC performance. The model achieves an L_{mse} of 0.0843 and L_{abs} of 0.2442, improving over $z_l + z_s + z_e$ with *statistical significance*. The SEC capabilities can be further noted in the audio examples presented online[1].

Influence of segment size: While training HiFiGAN, to account for varying utterance lengths, segments of consistent lengths are randomly sampled from different regions of the full utterance and then are used for training. In [22], it is noted that the HiFiGAN is robust enough to resynthesise natural-sounding speech even when trained on a small segment size (0.75s). However, for modeling emotion the context is important [34], thereby requiring larger segment sizes to aptly condition on the emotion content.

To investigate the influence of segment sizes on SEC, we trained the proposed methodology on varying segment sizes, and the results are presented in Fig. 2. From the results, it is noted that larger segment sizes improve SEC performances, both in terms of L_{mse} and WVMOS. While the improvements on naturalness (WVMOS) are clearly noted, improvements on L_{mse} are larger for extreme emotion values (1 and 7) than mid-ranges of emotion (3 to 5).

Performance on different arousal classes: Existing research has shown that SEC techniques generally tend to perform well on certain emotion pairs than others. For example, in [17], it is noted that the emotion pair of angry-sad is easier to convert than the happy-angry pair. To investigate this, Fig. 2 also present performances with respect to each of the arousal classes ranging from 1 to 7. In terms of the SEC capability, the L_{mse} error is noted to be higher for extreme arousal classes (1 and 7). However, for mid-scale arousal values between 2 and 6 the L_{mse} error is comparatively smaller. In terms of the naturalness of synthesised speech, the WVMOS score is lower for high arousal classes (6 and 7) and is comparatively better for other arousal classes (1 to 5). This reveals that the proposed methodology is capable of synthesising more natural-sounding speech when *reducing* the arousal of the input speech, than when *increasing* the arousal. Overall, the results reveal that the methodology better synthesises speech for mid-scale arousal than for extreme arousal.

Qualitative analysis of spectrograms: Sample spectrograms of emotion converted speech can be seen in Fig. 3, for the ground-truth speech of arousal $e = 3.20$, for synthesised speech of *reduced* arousal $\bar{e} = 1$, and for *increased* arousal $\bar{e} = 7$. Comparing the spectrograms of arousal 1 and arousal 7, from the marked eclipses, we can observe that for increased arousal of 7 the spectrograms have larger magnitudes in the mid-frequencies. This reveals that the model associates larger frequency magnitudes for high arousal speech, than for low arousal speech. Along with the spectrograms, we also plot the pitch contours of the respective speech signals. From the pitch contours, it can be noted that the synthesised speech for high arousal ($\bar{e} = 7$) has a higher mean and variability of pitch, than that of both the ground-truth speech ($e = 3.20$) and the synthesised speech for low arousal ($\bar{e} = 1$). This aligns with existing literature that associates high intensity of emotion with an increased mean pitch [14]. This difference in pitch is also clearly notable in the audio examples available online[1]. These results validate that the proposed model successfully performs SEC by aptly conditioning on the emotion content.

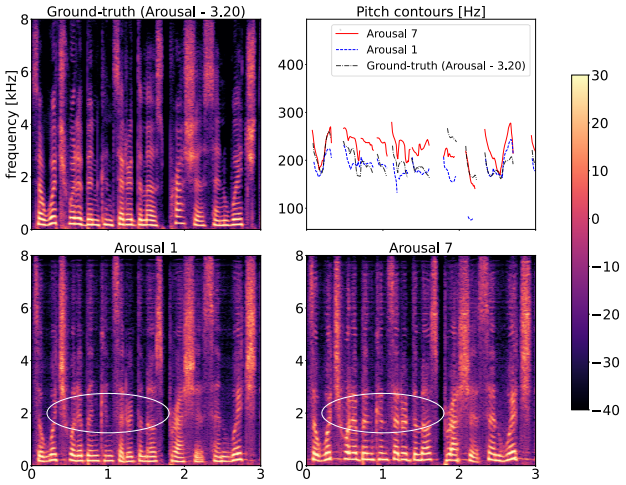


Figure 3: Sample log-energy spectrogram of emotion converted speech, along with comparisons on pitch contours.

6 Conclusion

In this paper, as the first in literature, we tackle the problem of in-the-wild speech emotion conversion on continuous arousal representations, as opposed to acted speech and categorical representations. In-the-wild datasets lack parallel utterances and thereby the problem of disentangling the lexical, speaker, and emotion information arises. To tackle this, we introduced a novel methodology that uses SSL encoders for disentanglement, and a HiFiGAN vocoder to resynthesise emotion conditioned speech from the disentangled SSL representations. We validated the network on an in-the-wild dataset, in terms of its emotion conversion capability, using a pretrained SER system, and the naturalness of synthesised speech, using WVMOS. Results reveal that the network is capable of synthesising natural-sounding speech with emotion conversion. Further analysis revealed that the network better synthesises emotional speech for mid-scale arousal than for extreme arousal. Finally, the pitch contour analysis showed that the synthesised speech for high arousal has a higher mean and variability of pitch than that of both the ground-truth speech and the synthesised speech for low arousal.

[1] https://uhh.de/inf-sp-emovc-itg

References

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] J. Crumpton and C. L. Bethel, "A survey of using vocal prosody to convey emotion in robot speech," *International Journal of Social Robotics*, vol. 8, pp. 271–285, 2016.

[3] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech synthesis with mixed emotions," *IEEE Tran. on Affective Computing*, 2022.

[4] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Inf. Proc. Sys., NeurIPS*, vol. 33, 2020.

[5] W.-N. Hsu, T. Remez, B. Shi, J. Donley, and Y. Adi, "Revise: Self-supervised speech resynthesis with visual input for universal and generalized speech regeneration," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023.

[6] A. Triantafyllopoulos, B. W. Schuller, G. İymen, M. Sezgin, X. He, Z. Yang, P. Tzirakis, S. Liu, S. Mertes, E. André, *et al.*, "An overview of affective speech synthesis and conversion in the deep learning era," *Proc. of the IEEE*, 2023.

[7] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[8] Z. Du, B. Sisman, K. Zhou, and H. Li, "Disentanglement of Emotional Style and Speaker Identity for Expressive Voice Conversion," in *Interspeech*, 2022.

[9] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[10] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[11] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[12] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The msp-conversation corpus," *Interspeech*, 2020.

[13] N. Raj Prabhu, G. Carbajal, N. Lehmann-Willenbrock, and T. Gerkmann, "End-to-end label uncertainty modeling for speech-based arousal recognition using Bayesian neural networks," in *Interspeech*, Sep 2022.

[14] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Tran. on Affective Computing*, vol. 14, no. 1, pp. 31–48, 2023.

[15] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, pp. 920–924, IEEE, 2021.

[16] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, *et al.*, "A database of german emotional speech.," in *Interspeech*, vol. 5, pp. 1517–1520, 2005.

[17] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, 2020.

[18] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," in *The Speaker and Language Recognition Workshop (Speaker Odyssey)*, 05 2020.

[19] K. Zhou, B. Sisman, and H. Li, "Vaw-gan for disentanglement and recomposition of emotional elements in speech," in *IEEE Spoken Language Technology Workshop*, 2021.

[20] D. de Oliveira, N. Raj Prabhu, and T. Gerkmann, "Leveraging semantic information for efficient self-supervised emotion recognition with audio-textual distilled models," in *Interspeech*, 2023.

[21] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 2023.

[22] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Interspeech*, 2021.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Inf. Proc. Sys., NeurIPS*, vol. 33, pp. 12449–12460, 2020.

[24] F. Kreuk, J. Keshet, and Y. Adi, "Self-Supervised Contrastive Learning for Unsupervised Phoneme Segmentation," in *Interspeech*, pp. 3700–3704, 2020.

[25] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.

[26] A. Sicherman and Y. Adi, "Analysing discrete self supervised speech representation for spoken language modeling," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, pp. 1–5, 2023.

[27] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, pp. 125–125, 2016.

[28] F. Kreuk, A. Polyak, J. Copet, E. Kharitonov, T. A. Nguyen, M. Rivière, W.-N. Hsu, A. Mohamed, E. Dupoux, and Y. Adi, "Textless speech emotion conversion using discrete & decomposed representations," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 2022.

[29] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Int. Conf. Machine Learning (ICML)*, 2016.

[30] L. I.-K. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, vol. 45, p. 255, Mar. 1989.

[31] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: A unified framework for bandwidth extension and speech enhancement," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, pp. 1–5, 2023.

[32] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z.-H. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Interspeech*, Apr 2018.

[33] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Tran. on Affective Computing*, vol. 7, pp. 190–202, Apr 2016.

[34] G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *IEEE Spoken Language Technology Workshop*, pp. 126–131, 2018.

4.2 Speech Emotion Conversion using Diffusion Models [P10]

Abstract

Speech emotion conversion is the task of converting the expressed emotion of a spoken utterance to a target emotion while preserving the lexical content and speaker identity. While most existing works in speech emotion conversion rely on acted-out datasets and parallel data samples, in this work we specifically focus on more challenging in-the-wild scenarios and do not rely on parallel data. To this end, we propose a diffusion-based generative model for speech emotion conversion, the EMOCONV-Diff, that is trained to reconstruct an input utterance while also conditioning on its emotion. Subsequently, at inference, a target emotion embedding is employed to convert the emotion of the input utterance to the given target emotion. As opposed to performing emotion conversion on categorical representations, we use a continuous arousal dimension to represent emotions while also achieving intensity control. We validate the proposed methodology on a large in-the-wild dataset, the MSP-Podcast v1.10. Our results show that the proposed diffusion model is indeed capable of synthesizing speech with a controllable target emotion. Crucially, the proposed approach shows improved performance along the extreme values of arousal and thereby addresses a common challenge in the speech emotion conversion literature.

Reference

N. Raj Prabhu and B. Lay and S. Welker and N. Lehmann-Willenbrock and T. Gerkmann, "EMOCONV-Diff: Diffusion-Based Speech Emotion Conversion for Non-Parallel and in-the-Wild Data", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Republic of Korea, April, 2024, pp. 11651-11655, DOI: 10.1109/ICASSP48485.2024.10447372.

Copyright Notice

Authors' Contributions

Navin Raj Prabhu led the study, including the initial conceptualization, algorithm development, neural network training, experimental validation, and manuscript preparation. Bunlong Lay and Simon Welker contributed through discussions on the methodological model design, provided insights into the experimental validation, and participated in the manuscript review. Nale Lehmann-Willenbrock contributed by reviewing the manuscript and helping to refine the argumentation and overall framing. Timo Gerkmann provided key insights into the experimental validation, offered valuable methodological feedback through discussions, and participated in the manuscript review.

EMOCONV-DIFF: DIFFUSION-BASED SPEECH EMOTION CONVERSION FOR NON-PARALLEL AND IN-THE-WILD DATA

Navin Raj Prabhu[⋆†] *Bunlong Lay*[⋆] *Simon Welker*[⋆] *Nale Lehmann-Willenbrock*[†] *Timo Gerkmann*[⋆]

[⋆]Signal Processing, Universität Hamburg, Germany
[†]Industrial and Organizational Psychology, Universität Hamburg, Germany
navin.raj.prabhu@uni-hamburg.de

ABSTRACT

Speech emotion conversion is the task of converting the expressed emotion of a spoken utterance to a target emotion while preserving the lexical content and speaker identity. While most existing works in speech emotion conversion rely on acted-out datasets and parallel data samples, in this work we specifically focus on more challenging in-the-wild scenarios and do not rely on parallel data. To this end, we propose a diffusion-based generative model for speech emotion conversion, the EmoConv-Diff, that is trained to reconstruct an input utterance while also conditioning on its emotion. Subsequently, at inference, a target emotion embedding is employed to convert the emotion of the input utterance to the given target emotion. As opposed to performing emotion conversion on categorical representations, we use a continuous arousal dimension to represent emotions while also achieving intensity control. We validate the proposed methodology on a large in-the-wild dataset, the MSP-Podcast v1.10. Our results show that the proposed diffusion model is indeed capable of synthesizing speech with a controllable target emotion. Crucially, the proposed approach shows improved performance along the extreme values of arousal and thereby addresses a common challenge in the speech emotion conversion literature.

Index Terms— Speech emotion conversion, diffusion models, non-parallel samples, arousal, in-the-wild

1. INTRODUCTION

Speech is one of the key social signals used by humans to express their emotions [1]. While significant developments have been made in speech generation and synthesis, *emotion-conditioned* speech synthesis is still a challenge [1, 2]. In the context of human-machine interaction, to improve the naturalness of machine communication, the generation of emotionally expressive speech is required [1]. Speech emotion conversion (SEC) is a sub-field of emotion-conditioned speech synthesis that aims to map a speech signal into another speech signal by converting its emotional expression while preserving the lexical information and the speaker's identity [3].

Emotions are represented in SEC as either *categorical* (e.g., six basic emotions [4]) [3, 5] or *continuous* (e.g., circumplex model [6]) [7] representations. It is well established in the speech emotion recognition (SER) and psychology literature that emotion is a complex construct with *fuzzy* class boundaries [8], and the categorical representations (e.g., happy, anger) do not aptly capture the subtle difference between human emotions [6]. The circumplex model contrarily represents emotions using *continuous* and independent dimensions, i.e., *arousal* (relaxed vs. activated) and *valence* (positive vs. negative) [6]. While the audio modality typically captures the arousal dimension of emotion well, it insufficiently explains valence [9, 10]. Therefore, in this work, we follow [7] and represent emotion using the continuous arousal dimension. Moreover, by using the continuous representations (arousal on a scale of 1 to 7) we directly achieve intensity control in SEC, as opposed to an additional effort required for categorical representations (e.g., [5, 11]).

Current SEC systems are typically trained on high-quality recorded speech data that are *acted-out* by professional actors. As a consequence the resulting algorithms are typically sensitive to noise and variabilities pertinent in real-world scenarios [12] (e.g., acoustic noise, speaker variabilities, subtle intonations, or vocal bursts that carry emotion; e.g., [13]). Furthermore, SEC systems trained on acted-out speech may create stereotypical portrayals of emotions [12]. Another crucial drawback of acted-out datasets is that they require *parallel* utterances, i.e., each source utterance is required to also have a ground-truth utterance of a target emotion [14, 15]. However, parallel utterances are expensive to collect [14], and models trained on them lack scalability [1]. In this work, we address these drawbacks of acted-out and parallel data by specifically focusing on non-parallel *in-the-wild* data.

A challenge in overcoming the usage of parallel utterances is the problem of *disentanglement*, where a disentanglement technique is required to decompose the source utterance into several constituents (i.e., emotion, lexical, and speaker information) before synthesizing speech for a target emotion [1, 3]. Existing works have employed encoder-decoders [5], generative adversarial networks [14], and self-supervised learning (SSL) [7] for the disentanglement. Recently, the so-called *diffusion models* have been introduced for the synthesis of high-quality samples, both in the audio- and image-domain [16, 17]. Further in [17], the disentanglement capability of diffusion models was uncovered for the task of text-conditioned image editing and demonstrated strong control over the image synthesis process.

For in-the-wild SEC without relying on parallel utterances, we introduce a diffusion-based approach that is trained to reconstruct a source utterance while also conditioning on its emotion. Subsequently, at inference, a target emotion embedding is employed to convert the emotion of the source utterance to the given target emotion. As such, the contributions of this paper are as follows: We introduce a novel emotion-conditioned diffusion model that does not rely on parallel utterances for SEC, which is in contrast to existing emotion-conditioned diffusion models that rely on parallel utterances and operate on the text-to-speech (TTS) domain [18, 11]. Building up on our previous work [7], our models can cope with unseen real-world scenarios, as it is trained on non-parallel in-the-wild

This work was funded under the Excellence Strategy of the Federal Government and the Länder, and the "Mechanisms of Change in Dynamic Social Interaction" project (LFF-FV79, Landesforschungsförderung Hamburg).

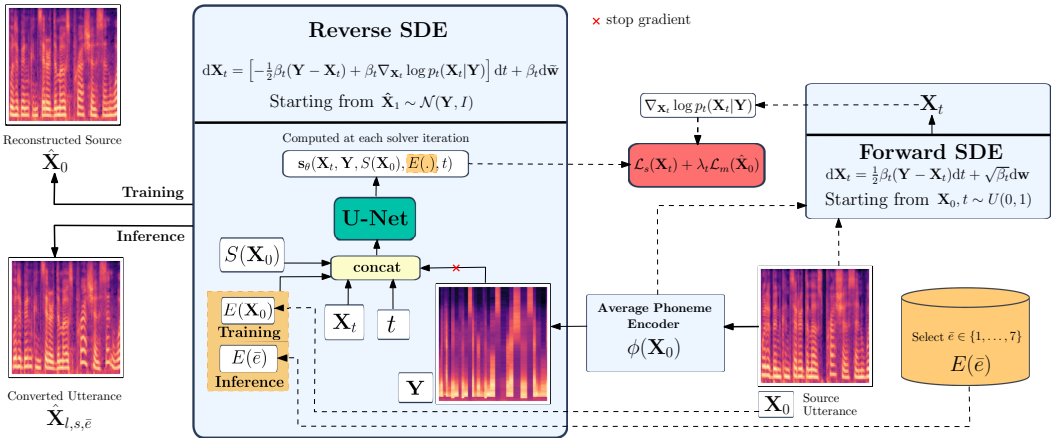


Fig. 1: Illustration of the training and inference process of the proposed EmoConv-Diff approach. Dotted arrows denote operations performed only during training. The *stop gradient* function stops the accumulation of the gradients of the inputs during the training.

speech utterances. To the best of our knowledge, we are the first to tackle this problem of non-parallel in-the-wild data for SEC, and the paper at hand is the first to employ diffusion models for this. Finally, the proposed approach improves over the *HiFiGAN* [7] for extreme target emotions, a common problem in SEC and TTS [7, 19].

2. DIFFUSION MODELS

Diffusion models are used in various applications across domains for the task of generation, such as image editing [17], speech enhancement [20], and TTS [21]. The idea behind these models involves adding Gaussian noise to the data using a stochastic differential equation (SDE). The *forward SDE* or *forward process* can be viewed as transforming an initial distribution into a terminating distribution that is usually tractable and available during inference. Under mild constraints, a forward SDE can be inverted by the *reverse SDE* [22]. The reverse SDE or *reverse process* transforms the terminating distribution of the forward process back into the initial distribution, during which the disentanglement is achieved [17].

In the extant literature, emotion-conditioned diffusion models rely on parallel data and operate on the TTS domain [11, 18]. In [18], the GradTTS-based *EmoDiff* was introduced. EmoDiff achieves emotion-conditioned speech synthesis from source text using a soft-label guidance technique in the reverse process. [11] introduces *EmoMix*, which uses pretrained SER embeddings of a reference utterance to exemplify the target emotional prosody and condition on the desired emotion. Note that both [18] and [11] rely on acted-out parallel utterances and operate on the TTS domain.

3. PROPOSED METHODOLOGY: EMOCONV-DIFF

We define the SEC task as follows: given the mel spectrogram of a source speech utterance $\mathbf{X}_{l,s,e}$ (or simply $\mathbf{X}_0$), containing lexical content l, speaker identity s, and emotion information e, we aim to generate a new mel spectrogram $\hat{\mathbf{X}}_{l,s,\bar{e}}$ that only transforms the arousal information to a target value $\bar{e}$. For this, we introduce a diffusion-based approach, the *EmoConv-Diff*, which is summarized in Fig. 1. The EmoConv-Diff comprises a set of *encoders*, each encoding the attributes to be disentangled, and a diffusion-based *decoder*, which aims to disentangle the attributes and perform emotion-controllable speech synthesis. The output of the diffusion decoder is a mel spectrogram $\hat{\mathbf{X}}_{l,s,\bar{e}} \in \mathbb{R}^{n \times T}$ and it is converted into time domain speech signal using a pretrained HiFiGAN vocoder [23].

3.1. Encoders

The EmoConv-Diff comprises three encoders: the *phoneme encoder* $\phi(.)$, the *speaker encoder* $S(.)$, and the *emotion encoder* $E(.)$.
Phoneme Encoding: Speaker- and emotion-independent "average voice" phoneme-level mel features are used to encode the lexical content l. Let $\mathbf{Y} := \phi(\mathbf{X}_0)$ be the "average voice" representation of the source audio, where $\phi(.)$ is the pretrained phoneme encoder. The transformer-based encoder, adopted from [24], has been used previously in voice conversion tasks. The encoder output (see $\mathbf{Y}$ in Fig. 1) has the same dimensions as the source mel $\mathbf{X}_0 \in \mathbb{R}^{n \times T}$.
Speaker Encoding: To encode the speaker identity, we use a pretrained speaker verification model $S(.)$ [25], following [24]. The output of $S(.)$ is a $d-vector$ speaker representation $S(.) \in \mathbb{R}^{128}$.
Emotion Encoding: To encode emotional information, we use a pretrained SSL-based SER system $E(.) \in \mathbb{R}^{1024}$, introduced in [26]. The $E(.)$ network was built by fine-tuning the Wav2Vec2-Large-Robust network [26] on the MSP-Podcast (v1.7) dataset [8].

3.2. Diffusion-based decoder

The diffusion-based decoder follows the SDE formalism by [21]. Specifically, let t be the continuous diffusion time-step variable describing the progress of the diffusion process. For $0 \leq t \leq 1$ the forward SDE of this work is given by

$$\mathrm{d}\mathbf{X}_t = \frac{1}{2}\beta_t(\mathbf{Y} - \mathbf{X}_t)\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\mathbf{w}, \tag{1}$$

where $\mathbf{w}$ is the standard Wiener process [27], $\mathbf{X}_t$ is the current process state with initial condition $\mathbf{X}_0 = \mathbf{X}_{l,s,e}$ and β_t is a non-negative function called the noise schedule. The process state $\mathbf{X}_t$ follows a Gaussian distribution [27, Section 5] that is called the *perturbation kernel*:

$$p_{0t}(\mathbf{X}_t|\mathbf{X}_0, \mathbf{Y}) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{X}_t; \boldsymbol{\mu}(\mathbf{X}_0, \mathbf{Y}, t), \sigma(t)^2\mathbf{I}\right). \tag{2}$$

The mean evolution of $\mu(\mathbf{X}_0, \mathbf{Y}, t)$, or simply $\mu(t)$, is given by

$$\mu(\mathbf{X}_0, \mathbf{Y}, t) = \alpha_t \mathbf{X}_0 + (1 - \alpha_t)\,\mathbf{Y}, \tag{3}$$

where $\alpha_t = e^{-\frac{1}{2}\int_0^t \beta_s ds}$ and the variance evolution is given by

$$\sigma(t)^2 = \left(1 - \alpha_t^2\right)\mathbf{I} \tag{4}$$

We represent the closed-form of α_t as β_t and set $\beta_t = b_0 + t(b_1 - b_0)$ and chose $b_0, b_1 > 0$ such that $\alpha_1 \approx 0$. In this case, the mean evolution describes an interpolation starting at $t = 0$ at the distribution of source $\mathbf{X}_0$ and terminating approximately at the distribution of "average voice" phoneme features $\mathbf{Y}$ at $t = 1$. The forward SDE (1) has an associated reverse SDE [22]:

$$\mathrm{d}\mathbf{X}_t = \left[-\frac{1}{2}\beta_t(\mathbf{Y} - \mathbf{X}_t) + \beta_t \nabla_{\mathbf{X}_t} \log p_t(\mathbf{X}_t|\mathbf{Y})\right]\mathrm{d}t + \beta_t \mathrm{d}\widetilde{\mathbf{w}}\,, \tag{5}$$

where $\mathrm{d}\widetilde{\mathbf{w}}$ is a Wiener process going backward through the diffusion time-steps. Moreover, the reverse process follows the same trajectory as the forward process, i.e. the reverse SDE starts approximately with the distribution of average-voice and terminates for $t = 0$ into the distribution of source-targets.

A network called the *score model* $\mathbf{s}_\theta(\mathbf{X}_t, \mathbf{Y}, S(\mathbf{X}_0), E(\mathbf{X}_0), t)$, or simply $\mathbf{s}_\theta(\mathbf{X}_t, t)$, is *trained* to approximate the *score function* $\nabla_{\mathbf{X}_t} \log p_t(\mathbf{X}_t|\mathbf{Y})$, i.e., the gradients of log-density of noisy data $\mathbf{X}_t$. We use the U-Net architecture from [24] as the score model $\mathbf{s}_\theta$. With the trained $\mathbf{s}_\theta$, we can then use the reverse SDE to generate an estimate of the source target $\mathbf{X}_0$ from the "average voice" $\mathbf{Y}$ given speaker identity $S(\mathbf{X}_0)$ and emotion embeddings $E(\mathbf{X}_0)$. An intuition behind the reverse process is that the diffusion-based decoder is trained to reconstruct $\mathbf{X}_0$ while learning the disentanglement between the speech attributes l, s, and e. With this setup, we overcome the need for parallel data during the training process.

During *inference*, a target emotion embedding $E(\bar{e})$ is employed to convert the emotion of the source utterance to the given target emotion. The target emotion embedding $E(\bar{e})$ is defined as the *averaged* emotion embedding of a set reference utterance samples belonging to the emotion category $\bar{e}$, as

$$E(\bar{e}) \coloneqq \frac{1}{|A_p(\bar{e})|} \sum_{\mathbf{X}_0 \in A_p(\bar{e})} E(\mathbf{X}_0), \tag{6}$$

where the set of reference samples $A_p(\bar{e})$ is defined to be the top $p = 20\%$ samples belonging to the particular target arousal $\bar{e}$.

3.3. Loss functions

The score model is trained on the *score matching* loss [28] which aims to approximate the score function. The score matching loss for $\mathbf{X}_0$ at time t is formulated as

$$\mathcal{L}_s(\mathbf{X}_t) = \mathbb{E}_{\epsilon_t}\left[||\mathbf{s}_\theta(\mathbf{X}_t, t) + \sigma(t)^{-1}\epsilon_t||_2^2\right] \tag{7}$$

where $\mathbf{X}_t = \mu(t) + \sigma(t)\epsilon_t$ and ϵ_t is sampled from $\mathcal{N}(0, \sigma(t))$. In addition to $\mathcal{L}_s$, we follow [7, 11] to use a mel spectrogram reconstruction loss for better conditioning on emotion attributes. $\mathcal{L}_m$ measures the L_1-norm:

$$\mathcal{L}_m(\hat{\mathbf{X}}_0) = \sum_x ||\mathbf{X}_0 - \hat{\mathbf{X}}_0||_1, \tag{8}$$

where $\hat{\mathbf{X}}_0$ is the mel spectrogram of synthesized speech. Note here that during the training of the score model it is expensive to obtain $\hat{\mathbf{X}}_0$ which requires solving the full reverse SDE. For this, in contrast to [11], we utilize a single-step approximation of $\hat{\mathbf{X}}_0$ by only relying on $\mathbf{X}_t$, $\mathbf{s}_\theta$, and $\mathbf{Y}$, which are available during training. We use Tweedie's formula [29] to approximate $\hat{\mathbf{X}}_0$ as

$$\hat{\mathbf{X}}_0 = \frac{\hat{\mu}(t) - (1 - \alpha_t)\,\mathbf{Y}}{\alpha_t}\,. \tag{9}$$

	DNSMOS ↑		SER Error ↓	
	SIG	OVRL	$\mathcal{L}_{mse}$	$\mathcal{L}_{abs}$
HiFiGAN [7]	**3.21**	**2.79**	0.084	24%
$\mathcal{L}_s$	3.20	2.78	0.091	25%
$\mathcal{L}_s + \mathcal{L}_m(\mathbf{X}_t)$	3.08	2.62	0.121	34%
$\mathcal{L}_s + \mathcal{L}_m(\mathbf{X}_0)$	**3.21**	2.78	**0.072***	**21**%*

Table 1: Overall performance of model versions. * indicates statistically significant improvements in results.

where $\hat{\mu}(t)$ is an estimate of $\mu(t)$ (3), and is formulated as $\hat{\mu}(t) = \mathbf{X}_t - (\mathbf{s}_\theta(\mathbf{X}_t, t) * \sigma(t)^2)$. With that, the final loss function is

$$\mathcal{L}(\mathbf{X}_t, \hat{\mathbf{X}}_0) = \mathcal{L}_s(\mathbf{X}_t) + \lambda_t \mathcal{L}_m(\hat{\mathbf{X}}_0), \tag{10}$$

where λ_t is a weighting function depending on the current diffusion time-step t. Considering that $\mathbf{X}_t$ contains more Gaussian noise for larger t, we set $\lambda_t = 1 - t^2$, thereby weighting more for smaller t values and gradually decreasing the weights for larger t.

4. EXPERIMENTAL SETUP

Dataset: The proposed methodology is trained and validated on the *in-the-wild* MSP-Podcast dataset (v1.10) [8]. The dataset in contrast to predominant SEC datasets (e.g., ESD [30], IEMOCAP [31]) is larger (≈238hrs of audio), has utterances of variable duration, has over 1400 speakers, and contains naturalistic emotional expressions. For example, the ESD contains acted-out utterances from only 10 English speakers and only ≈29 hours of acted-out utterances. The arousal annotations, collected at the utterance-level on a scale of 1 to 7, are distributed with μ =4 and σ =0.95.

Validation measures: We validate the proposed methodology in terms of both the SEC capabilities and the speech quality of the synthesized signal. As the measure of SEC capability, we use the mean-squared L_{mse} and mean-absolute L_{abs} errors, calculated between the target arousal $\bar{e}$ and the SER prediction on the synthesized output $E(\hat{\mathbf{X}})$. As the measure of speech quality, we use the DNSMOS [32], a non-intrusive objective speech quality metric designed to predict the mean-opinion score (on a scale of 1 to 5) results of subjective listening tests (i.e., P.835 [32]). Specifically, we use the metric measuring the overall signal quality *OVRL*, and specifically the speech quality *SIG*. Note here that intrusive metrics cannot be used to evaluate in-the-wild recordings, like our dataset, as the reference is not available due to the lack of parallel data. Statistical significance for improved performance is estimated using one-tailed t-test on error distributions, asserting significance for p-values ≤ 0.05.

5. RESULTS AND DISCUSSION

Overall performance: We validate the overall performance of the proposed *EmoConv-Diff* against a baseline, the HiFiGAN-based SEC system [7], henceforth mentioned as *HiFiGAN*, which to the best of our knowledge is the only prior work on SEC using in-the-wild and non-parallel data. In addition to HiFiGAN [7], we use three different versions of the EmoConv-Diff, (i) $\mathcal{L}_s$, which is only trained on the score matching loss $\mathcal{L}_s$, (ii) $\mathcal{L}_s + \mathcal{L}_m(\mathbf{X}_t)$, which also uses the mel reconstruction loss $\mathcal{L}_m$ tuned on $\mathbf{X}_t$, and (iii) $\mathcal{L}_s + \mathcal{L}_m(\mathbf{X}_0)$, where the mel reconstruction loss $\mathcal{L}_m$ is tuned on the approximated source mel spectrogram $\hat{\mathbf{X}}_0$ (9).

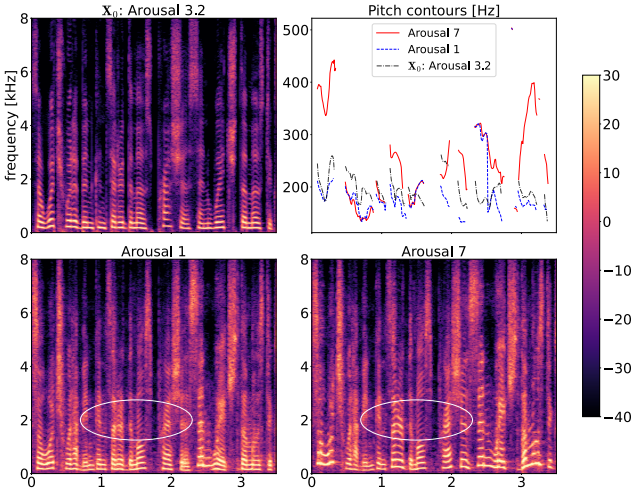


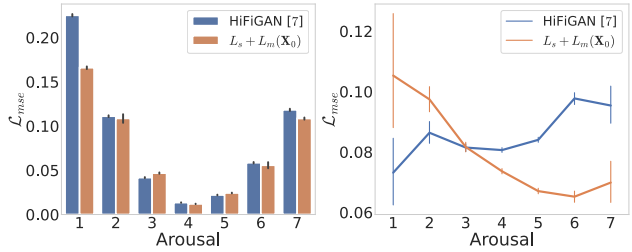
Fig. 2: Sample log-energy spectrogram of emotion converted speech, along with comparisons on pitch contours.

From the results presented in Table 1, we note the following. First, the EmoConv-Diff version $\mathcal{L}_s + \mathcal{L}_m(\mathbf{X}_0)$ achieves the best SER errors with statistical significance, achieving $\mathcal{L}_{mse}$ of 0.072 and $\mathcal{L}_{abs}$ of 21%. This confirms the emotion conversion capability of the proposed diffusion model. Second, in terms of the speech quality and overall signal quality, the EmoConv-Diff version $\mathcal{L}_s + \mathcal{L}_m(\mathbf{X}_0)$ performs on par with the HiFiGAN baseline. The variant achieves speech quality performance of 3.21 SIG and an overall signal quality of 2.78 OVRL. Third, the introduction of the mel reconstruction loss $\mathcal{L}_m(\hat{\mathbf{X}}_0)$ tuned on the derived approximation of source $\mathbf{X}_0$ (9) improves the performance of the diffusion model, in terms of both the DNSMOS scores and SER errors. Finally, when the mel reconstruction loss $\mathcal{L}_s$ is tuned on $\mathbf{X}_t$, the performance in terms of the SER errors diminishes, signifying the noisy nature of $\mathbf{X}_t$ and the efficiency of the derived $\hat{\mathbf{X}}_0$ during the training phase.

Qualitative analysis of spectrograms: Fig. 2 shows sample spectrograms of the source speech $\mathbf{X}_0$ of arousal $e = 3.20$, the converted speech of *reduced* arousal $\bar{e} = 1$, and of *increased* arousal $\bar{e} = 7$. A high average pitch of the speech signal is directly associated with a high intensity of emotion [5, 7]. Therefore, in Fig. 2, we also plot the pitch contours of the respective converted speech and the source $\mathbf{X}_0$. Comparing the spectrograms of arousal 1 and arousal 7, from the marked eclipses, we can observe that for increased arousal of 7 the spectrograms have larger magnitudes in the mid-frequencies. This reveals that the proposed EmoConv-Diff model associates *larger frequency magnitudes* for high arousal speech than for low arousal speech. From the pitch contours, it can be further noted that the synthesized speech for high arousal ($\bar{e} = 7$) has a *higher mean and variability of pitch*, than that of both the ground-truth speech ($e = 3.20$) and the synthesized speech for low arousal ($\bar{e} = 1$). This difference in pitch is also clearly notable in the audio examples available online[1]. These results show that the proposed model successfully performs SEC by aptly conditioning on the emotion content.

Performance for target arousal $\bar{e}$: SEC systems generally tend to perform well on certain emotion pairs and emotion classes. For example, [14] notes that the emotional pairing of "angry" and "sad" is easier to convert than the pairing of "happy" and "angry". Moreover,

[1] https://uhh.de/inf-sp-emoconvdiff



(a) Target arousal $\bar{e}$ (b) Source arousal e

Fig. 3: Class-wise L_{mse} performances for target arousal $\bar{e}$ and ground-truth arousal e.

given that the emotion classes are imbalanced in in-the-wild datasets [8], with fewer samples along the extremes of emotion scale, SEC for extreme values of $\bar{e}$ is a general challenge [7, 19]. To investigate this, in Fig. 3a, we plot the $\mathcal{L}_{mse}$ performance with respect to each of the target arousal classes $\bar{e}$. From the plot, we note that the proposed EmoConv-Diff model makes the *largest* improvements along the extreme target arousal values (i.e., $\bar{e}$ =1 and 7) while performing on par along the mid scale arousal values (i.e., $\bar{e}$ =2, 3, 4, 5 and 6). This confirms that the proposed EmoConv-Diff overcomes a crucial shortcoming of existing SEC systems by improving along the extreme values of $\bar{e}$.

Performance for source arousal e: While it is important to evaluate the SEC performance with respect to the target arousal $\bar{e}$, it is also important to the SEC performance with respect to the arousal of the source speech $\mathbf{X}_0$ (i.e., e). In Fig. 3b, we also plot the $\mathcal{L}_{mse}$ performances with respect to e and observe the following. First, for both the HiFiGAN [7] and the proposed EmoConv-Diff, the standard deviation of $\mathcal{L}_{mse}$ with respect to the extreme emotions (e =1 and 7) is *larger* than the mid scale values of e. This indicates that it is generally harder for SEC systems to convert the emotion of source $\mathbf{X}_0$ with already extreme emotions (i.e., e =1 and 7), while it is easier to convert the emotion neutral emotion source $\mathbf{X}_0$ (i.e., e =3, 4 and 5). Second, we note contrasting behaviors between the HiFiGAN and the EmoConv-Diff. While the EmoConv-Diff achieves better performance for *higher* source arousal values (e >3) than *lower* arousal values, the HiFiGAN does better for *lower* source arousal values (e <3) than *higher* arousal values. Moreover, the proposed EmoConv-Diff model performs better than the HiFiGAN in four of the seven arousal classes, which points to the superior SEC capability of the EmoConv-Diff compared to the HiFiGAN baseline.

6. CONCLUSION

Emotion-conditioned speech synthesis (ESS) is an important application that can promote the naturalness of machine communication. Speech emotion conversion (SEC) is a sub-field of ESS. In this paper, we moved beyond the typical reliance on acted-out data sets and parallel samples in SEC, by proposing a diffusion-based generative model and using the continuous arousal dimension to represent emotions while also achieving intensity control. We validated our model using the MSP-Podcast v1.10, a large in-the-wild dataset. We show that our proposed diffusion model, the EmoConv-Diff, is indeed able to synthesize speech for a controllable target emotion. In particular, in comparison to our prior work [7], our model shows improved performance along the extreme values of arousal and thereby addresses a common challenge in the SEC literature [7, 19].

7. REFERENCES

[1] A. Triantafyllopoulos, B. W. Schuller, G. İymen, M. Sezgin, X. He, Z. Yang, P. Tzirakis, S. Liu, S. Mertes, E. André, *et al.*, "An overview of affective speech synthesis and conversion in the deep learning era," *Proc. of the IEEE*, 2023.

[2] S. Amiriparian, B. W. Schuller, N. Asghar, H. Zen, and F. Burkhardt, "Guest editorial: Special issue on affective speech and language synthesis, generation, and conversion," *IEEE Tran. on Affective Computing*, 2023.

[3] Z. Du, B. Sisman, K. Zhou, and H. Li, "Disentanglement of Emotional Style and Speaker Identity for Expressive Voice Conversion," in *Interspeech*, Sep 2022.

[4] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[5] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Tran. on Affective Computing*, 2023.

[6] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, 1980.

[7] N. Raj Prabhu, N. Lehmann-Willenbrock, and T. Gerkmann, "In-the-wild speech emotion conversion using disentangled self-supervised representations and neural vocoder-based resynthesis," in *Proc. ITG Conf. on Speech Comm.*, Sept. 2023.

[8] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Tran. on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[9] D. de Oliveira, N. Raj Prabhu, and T. Gerkmann, "Leveraging semantic information for efficient self-supervised emotion recognition with audio-textual distilled models," in *Interspeech*, 2023.

[10] N. Raj Prabhu, G. Carbajal, N. Lehmann-Willenbrock, and T. Gerkmann, "End-to-end label uncertainty modeling for speech-based arousal recognition using Bayesian neural networks," in *Interspeech*, Sep 2022.

[11] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "EmoMix: Emotion Mixing via Diffusion Models for Emotional Speech Synthesis," in *Interspeech*, 2023.

[12] K. Zhou, *Emotion modelling for speech generation*. Phd thesis, National University of Singapore, 2022. Available at `https://scholarbank.nus.edu.sg/handle/10635/243782`.

[13] F. Busquet, F. Efthymiou, and C. Hildebrand, "Voice analytics in the wild: Validity and predictive accuracy of common audio-recording devices," *Behavior Research Methods*, 2023.

[14] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2020.

[15] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," in *The Speaker and Language Recognition Workshop (Speaker Odyssey)*, May 2020.

[16] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE Tran. on on Audio, Speech, and Language Processing*, 2023.

[17] Q. Wu, Y. Liu, H. Zhao, A. Kale, T. Bui, T. Yu, Z. Lin, Y. Zhang, and S. Chang, "Uncovering the disentanglement capability in text-to-image diffusion models," in *IEEE/CVF Conf. on Computer Vision and Pattern Rec. (CVPR)*, June 2023.

[18] Y. Guo, C. Du, X. Chen, and K. Yu, "Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2023.

[19] S. Wang, J. Guðnason, and D. Borth, "Learning Emotional Representations from Imbalanced Speech Data for Speech Emotion Recognition and Emotional Text-to-Speech," in *Interspeech*, 2023.

[20] B. Lay, S. Welker, J. Richter, and T. Gerkamnn, "Reducing the prior mismatch of stochastic differential equations for diffusion-based speech enhancement," *Interspeech*, 2023.

[21] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *Int. Conf. Machine Learning (ICML)*, PMLR, 2021.

[22] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, 1982.

[23] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, 2020.

[24] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *Int. Conf. on Learning Representations (ICLR)*, 2022.

[25] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu, *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 31, 2018.

[26] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Tran. on Pattern Analysis and Machine Int.*, 2023.

[27] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*. Springer, 2nd ed., 1996.

[28] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *Advances in Neural Inf. Proc. Systems (NeurIPS)*, 2020.

[29] B. Efron, "Tweedie's formula and selection bias," *Journal of the American Statistical Association*, vol. 106, no. 496, 2011.

[30] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[31] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, 2008.

[32] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2022.

4.3 Duration Modeling for Speech Emotion Conversion [P11]

Abstract

Speech Emotion Conversion aims to modify the emotion expressed in input speech while preserving lexical content and speaker identity. Recently, generative modeling approaches have shown promising results in changing local acoustic properties such as fundamental frequency, spectral envelope and energy, but often lack the ability to control the duration of sounds. To address this, we propose a duration modeling framework using resynthesis-based discrete content representations, enabling modification of speech duration to reflect target emotions and achieve controllable speech rates without using parallel data. Experimental results reveal that the inclusion of the proposed duration modeling framework significantly enhances emotional expressiveness, in the in-the-wild MSP-Podcast dataset. Analyses show that low-arousal emotions correlate with longer durations and slower speech rates, while high-arousal emotions produce shorter, faster speech.

Reference

N. Raj Prabhu and D. de Oliveira and N. Lehmann-Willenbrock and T. Gerkmann "Enhancing In-the-Wild Speech Emotion Conversion with Resynthesis-based Duration Modeling", *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Honolulu, Hawaii, USA, December, 2025.

Copyright Notice

Authors' Contributions

Navin Raj Prabhu led the study, including the initial conceptualization, algorithm development, neural network training, experimental validation, and manuscript preparation. Danilo de Oliveira contributed by providing feedback on experimental validation through discussions and assisted in reviewing the manuscript. Nale Lehmann-Willenbrock contributed by reviewing the manuscript and helping to refine the argumentation and overall framing. Timo Gerkmann provided key insights into the experimental validation, offered valuable methodological feedback through discussions, and participated in the manuscript review.

Enhancing In-the-Wild Speech Emotion Conversion with Resynthesis-based Duration Modeling

Navin Raj Prabhu, Danilo de Oliveira
Signal Processing
University of Hamburg
Hamburg, Germany
{firstname.lastname}@uni-hamburg.de

Nale Lehmann-Willenbrock
Industrial and Organizational Psychology
University of Hamburg
Hamburg, Germany
nale.lehmann-willenbrock@uni-hamburg.de

Timo Gerkmann
Signal Processing
University of Hamburg
Hamburg, Germany
timo.gerkmann@uni-hamburg.de

***Abstract*—Speech Emotion Conversion aims to modify the emotion expressed in input speech while preserving lexical content and speaker identity. Recently, generative modeling approaches have shown promising results in changing local acoustic properties such as fundamental frequency, spectral envelope and energy, but often lack the ability to control the duration of sounds. To address this, we propose a duration modeling framework using resynthesis-based discrete content representations, enabling modification of speech duration to reflect target emotions and achieve controllable speech rates without using parallel data. Experimental results reveal that the inclusion of the proposed duration modeling framework significantly enhances emotional expressiveness, in the in-the-wild MSP-Podcast dataset. Analyses show that low-arousal emotions correlate with longer durations and slower speech rates, while high-arousal emotions produce shorter, faster speech.**

***Index Terms*—Speech emotion conversion, duration modeling, non-parallel samples, arousal, in-the-wild**

I. Introduction

Speech is a fundamental social signal that plays a key role in enabling interactions, whether between humans or between humans and machines. It conveys essential information for the interaction, including lexical content, speaker identity, and expressed emotions [1]. The task of speech generation and synthesis thereby is a crucial research topic in the fields of signal processing and human-computer interaction. With the advent of generative deep neural networks, substantial improvements have been made in speech generation and synthesis [2]–[5]. However, emotion-conditioned speech synthesis remains a significant challenge [6]–[9]. In the context of human-computer interaction, the need for emotion-conditioned speech synthesis is crucial, not only to improve the naturalness and expressiveness of machine communication but also to enhance user engagement, foster empathy, and enable more effective and context-aware responses [9], [10].

Speech Emotion Conversion (SEC) is a sub-field of emotion-conditioned speech synthesis that aims to modify the emotion expressed in input speech while preserving lexical content and speaker identity [6], [8], [11]. This requires precise control over prosodic attributes that convey emotional content, such as intonation, stress, rhythm, and loudness, which are controlled by the acoustic features of speech sounds, such as fundamental frequency, duration, energy, and spectral envelope. While it is appealing to control these attributes based on a target emotion, changing the corresponding acoustic feature for each prosodic component presents its own unique set of challenges [9].

Generative deep neural networks, such as variational autoencoders (VAEs) [12], generative adversarial networks (GANs) [13], and diffusion models [14], have been employed to the task of SEC with success in emotion conversion capabilities and improved naturalness in generated speech [10], [15]. However, these methods often overlook duration modeling in emotion conversion, resulting in inadequate control over crucial prosodic features such as rhythm and stress. Instead, they typically enforce fixed durations, where the emotion-converted output speech sounds have exactly the same duration as in the input, regardless of the intended emotion change. Interestingly, this is in contrast to the task of text-to-speech (TTS) synthesis, where duration modeling with duration-flexible speech generation is a common module, with proven improvements in the naturalness of synthesised speech [4].

Durflex-EVC [9] introduced duration modeling in SEC with parallel data, where for each source utterance with a corresponding source emotion also a corresponding target utterance with a target emotion is available. Durflex-EVC learns discrete speech units from parallel target emotion speech and their repetitions. However, a particular challenge in duration modeling for emotion conversion arises when working with in-the-wild emotion datasets, as these lack parallel samples. As a result, there is no ground-truth duration reference for the target emotion, making accurate duration control more challenging. While in-the-wild datasets offer a richer and more naturalistic collection of emotional speech, along with greater speaker diversity and varied acoustic conditions [16]–[18], their non-parallel nature limits their applicability for supervised duration modeling in emotion conversion [10], [15]. In this work, we aim to achieve duration modeling in SEC, focusing specifically on in-the-wild datasets without relying on parallel data.

In this work, we propose a resynthesis-based duration modeling approach to enhance SEC performance, which operates on discrete speech units and does not require parallel data.

This work was funded under the Excellence Strategy of the Federal Government and the Länder, and the project "Mechanisms of Change in Dynamic Social Interaction" (LFF-FV79, Landesforschungsförderung Hamburg).

To enable duration modeling in a non-parallel setting, the proposed method is trained using a resynthesis setup, inspired by [10] and [15]. In this setup, during training, the model simultaneously reconstructs the original input speech while the duration model learns to predict the repetition counts of discrete speech units. This prediction is based solely on the input speech, without any reliance on target speech. During inference, the trained duration model can predict the appropriate unit repetitions based on a target emotion embedding, enabling emotion-aware duration control. Experiments in the in-the-wild MSP-Podcast dataset show that the inclusion of the proposed duration modeling framework is beneficial for emotional expressiveness.

II. Related Literature

A. Speech Emotion Conversion techniques

SEC techniques can be broadly categorized into *sequential* speech generation and *parallel* speech generation models. Sequential generation models (e.g., [19]–[21]) perform emotion-conditioned speech synthesis by sequentially generating speech units or frames, thereby achieving implicit duration modeling. However, they often face challenges such as difficulty in capturing long-term dependencies and high time complexity [9]. This has motivated the development of parallel generation models (e.g., [7], [9], [22]), which address these limitations by enabling parallel generation of speech frames. However, a key requirement of these models is the explicit modeling of the intended duration [9].

Recently, there has been a shift in voice and emotion conversion research away from traditional scripted or acted-out speech, which often lacks the natural spontaneity of real-life conversations, towards the use of in-the-wild recorded speech [10], [15], [16]. Unlike acted-out data, which is essentially read-out speech, in-the-wild recordings are more spontaneous and capture diverse speaking styles, emotional expressions, nonverbal cues like laughter and lip smacks, and disfluencies such as repetitions, hesitations, and interruptions [16], [23]. Empirical analyses using the NaturalVoices dataset [16] show that models trained on in-the-wild samples generate more natural and intelligible speech. However, such training requires methods that do not depend on parallel speech samples.

Raj Prabhu et al. [10] proposed a SEC framework using *resynthesis* to eliminate the need for parallel data. A HiFiGAN-based vocoder reconstructs input speech from disentangled self-supervised learning (SSL)-based representations: discrete HuBERT embeddings for lexical content, speech emotion recognition (SER)-derived emotion embeddings, and speaker verification-based speaker embeddings. At inference, modifying the emotion embedding enables synthesis with the target emotion. Building on this, EmoConv-Diff [15] uses a diffusion decoder conditioned on "average-phoneme" mel features. While effective for in-the-wild SEC, these approaches lack duration modeling and cannot control speech rate based on the target emotion.

In this work, we use the resynthesis technique to achieve duration modeling under in-the-wild conditions without relying on any information from target emotion speech, neither target emotion durations nor speech embeddings. To the best of our knowledge, this is the first study to propose duration-flexible SEC that does not rely on parallel emotion speech samples.

B. Duration Modeling techniques

Duration Modeling in speech synthesis has been approached as a task of predicting the temporal alignment between lexical tokens (e.g., phonemes) and their respective acoustic features, essentially determining how long each unit should be held in the synthesised speech [4], [24]. Modern neural TTS systems incorporate duration modeling either implicitly, using the attention mechanism [25], [26], or explicitly, using a duration predictor that predicts phoneme repetitions [4], [24], [27]. The explicit modeling approach has been preferred in non-autoregressive models like Grad-TTS [4] and FastSpeech [24], allowing for greater flexibility in modifying speaking style, emphasis, or speech rate.

Despite its demonstrated effectiveness in TTS, duration modeling has received limited attention in tasks like emotion and voice conversion. A likely reason for this omission is the difficulty of jointly training a duration model and learning to modify the prosodic features of the input speech during conversion. As a result, many emotion conversion models adopt a fixed-duration strategy, where the converted speech maintains the same duration as the input, regardless of the target emotion [10], [15]. This constraint limits the expressiveness of SEC systems by restricting their ability to adjust the timing of lexical units, and consequently, the rhythm and speech rate aligned with the intended emotional state.

DurFlex-EVC [9] addresses the gap of incorporating duration modeling in SEC by using a so-called *Unit Aligner* module to extract discrete content tokens and a *Duration Predictor* to estimate their repetitions. However, this approach is not directly applicable to in-the-wild datasets, as it relies on speech units extracted from parallel target speech, which are unavailable in non-parallel settings. Additionally, the use of look-up table-based speaker and emotion embeddings further limits its adaptability to in-the-wild scenarios, where speaker and emotion conditions are more variable and less structured. Similarly, [28] and [29] also address duration modeling. While [28] target speaker conversion and [29] focus on emotion conversion with acted data and categorical emotion labels.

In this work, inspired by [10] and [15], we propose a resynthesis-based duration modeling approach that is better suited for in-the-wild datasets. Our method relies solely on the input speech during training and does not require any information from the target emotion or target speech, making it fully compatible with non-parallel SEC tasks.

III. Methodology

The overall task of speech emotion conversion can be formulated as follows: given a single-channel audio input $\mathbf{x}_{l,s,e} \in \mathbb{R}^{1\times T}$ representing a spoken utterance with lexical content l, speaker identity s, and annotated emotion level e, where the raw waveform is denoted as a sequence of samples

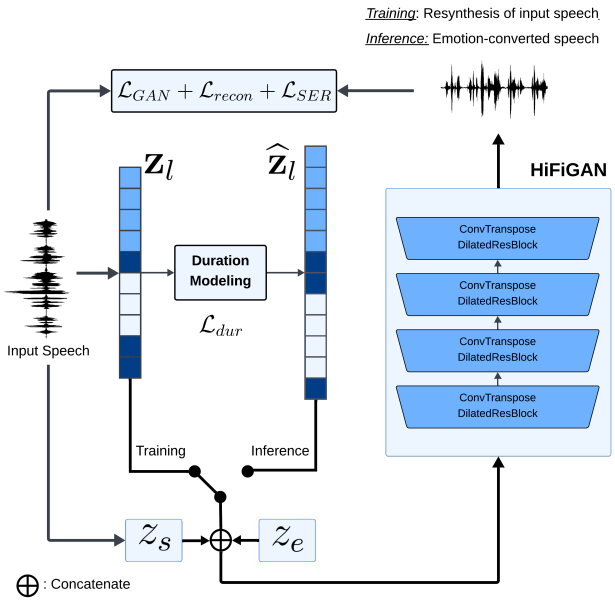


Fig. 1: Overview of the speech emotion conversion framework.

$\mathbf{x} = [x_1, ..., x_T]$, the goal is to generate $\widehat{\mathbf{y}}_{l,s,\bar{e}} \in \mathbb{R}^{1\times T'}$, with T' potentially different from T. This output should preserve the original lexical content l and speaker identity s from $\mathbf{x}_{l,s,e}$, while converting the expressed emotion to a desired target level $\bar{e}$. With that intent, the length T' is jointly modeled and the generated output is expected to be duration-flexible with respect to the desired target emotion $\bar{e}$. We adopt the SSL-based HiFiGAN model from [10] as the SEC backbone for integrating our resynthesis-based duration modeling approach. Its original design, which is also trained using a resynthesis paradigm, makes it particularly well-suited for this purpose. The overall SEC methodology is depicted in Figure 1.

A. Disentangled Representations

For the disentangled SSL-based representations input to the HiFiGAN decoder, we use the following encoded features:

(i) *Lexical representation* ($\mathbf{z}_l \in \mathbb{N}^{1\times N}$): Following [9]–[11], we use discrete HuBERT units obtained via k-means clustering on continuous HuBERT features. Formally, $\mathbf{z}_l = [z_1, ..., z_N]$, where each z_i is a positive integer and N is the length of the input discrete unit sequence, corresponding to the number of frames in HuBERT's representations. Prior studies [30]–[32] have shown that these units strongly correlate with the phonemic content of the utterance. The feature rate of these speech units is 49Hz.

(ii) *Speaker representation* ($z_s \in \mathbb{R}^{512}$): Adopted from [3], we use a d-vector extracted from a pretrained WavLM-based speaker verification model [33].

(iii) *Emotion representation* ($z_e \in \mathbb{R}^{128}$): A continuous embedding obtained by applying a trainable linear transformation to the emotion label e during training, and to the target emotion label $\bar{e}$ during inference.

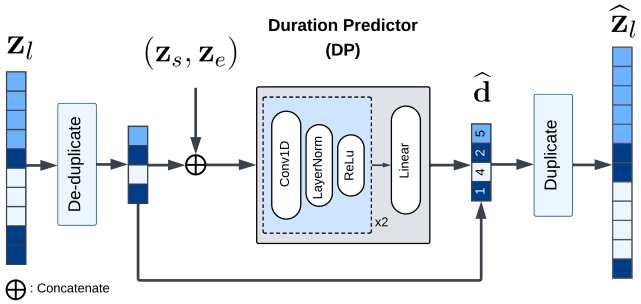


Fig. 2: Overview of the duration modeling technique.

Unlike $\mathbf{z}_l$, both z_s and z_e are global utterance-level representations. To align them with the frame-level $\mathbf{z}_l$, we broadcast z_s and z_e across frames/discrete units, resulting in $\mathbf{z}_s$ and $\mathbf{z}_e$.

B. Duration Modeling

An overview of the duration modeling technique can be seen Figure 2. Based on $\mathbf{z}_s$ and $\mathbf{z}_e$, we perform duration modeling on the discrete HuBERT speech units $\mathbf{z}_l$, which represent the lexical content of input speech. Formally, we formulate the resynthesis-based duration modeling as follows: for $\mathbf{z}_l$ of input speech, we train a *Duration Predictor* (DP) to predict the consecutive repetition of discrete speech units $\mathbf{d}$, conditioned on emotion and speaker representations. These repetitions represent the durations of each lexical unit.

Firstly, the frame-level $\mathbf{z}_l$ is de-duplicated to extract unit-level speech units, where the repetitions are ignored to obtain consecutive unique speech units. Secondly, this unit-level representation is fed as input to the predictor DP. To further achieve speaker and emotion conditioned duration modeling, we concatenate the speaker and emotion embeddings $(\mathbf{z}_s, \mathbf{z}_e)$ and pass them as an additional input to the predictor.

The predictor is a simple deterministic neural network comprising two convolution layers and a linear layer to predict $\mathbf{d}_i$ for respective unit-level speech units, where i is the index in the de-duplicated sequence of discrete speech units. As an example, if the speech units in $\mathbf{z}_l$ are $[1, 1, 2, 2, 2, 1, 3, 3, 3, 3]$, the de-duplicated sequence would be $[1, 2, 1, 3]$, and the target $\mathbf{d}$ would be $[2, 3, 1, 4]$. For stable training and to better account for outliers in durations, we predict durations in the log-scale: $\log(\mathbf{d})$, as suggested in [4]. During training, the predicted log-scale durations/repetitions are directly used in the loss function and the true frame-level $\mathbf{z}_l$ is used as the input to the HiFiGAN decoder. However, during inference, the predicted log durations $\widehat{\log(\mathbf{d})}$ are reversed back into duration units as follows:

$$\widehat{\mathbf{d}} = \min\left(1, e^{\widehat{\log(\mathbf{d})}+1}\right). \quad (1)$$

Finally, the reversed durations $\widehat{\mathbf{d}}$ are used to duplicate the unit-level speech units to obtain the duration modeled discrete lexical units $\widehat{\mathbf{z}}_l$. Note that, as per the resynthesis paradigm, we use the estimated $\widehat{\mathbf{z}}_l$ only during inference, and during training the true $\mathbf{z}_l$ is used. The final input to the HiFiGAN decoder is

the combined concatenated representation: $(\mathbf{z}_l, \mathbf{z}_s, \mathbf{z}_e)$ during training and $(\widehat{\mathbf{z}}_l, \mathbf{z}_s, \mathbf{z}_e)$ at inference time.

C. Loss Functions

The overall training of the SEC architecture involves four different loss terms: (i) the adversarial based HiFiGAN loss $\mathcal{L}_{GAN}$, which is the same as used in [11] and [10], (ii) a reconstruction loss,

$$\mathcal{L}_{recon}(G) = \sum_{\mathbf{x}} ||\phi(\mathbf{x}) - \phi(\hat{\mathbf{y}})||_1, \quad (2)$$

where ϕ is a function computing Mel-spectrogram, (iii) a speech emotion recognition loss which is used to better condition the SEC model on the emotion of input speech,

$$\mathcal{L}_{SER} = \sum_{\mathbf{x}} [1 - L_{ccc}(e, E_{SER}(\hat{\mathbf{y}}))], \quad (3)$$

where L_{ccc} is the concordance correlation coefficient (CCC) [34] computed between the ground-truth emotion e of input speech, and the predicted emotion for resynthesised speech $E_{SER}(\hat{\mathbf{y}})$, and finally, (iv) the duration modeling loss $\mathcal{L}_{dur}$. We use the speech emotion recognition (SER) model introduced in [35] as the emotion predictor $E_{SER}(.)$. The emotion predictor is a wav2vec-based neural network trained on the MSP-Podcast dataset to predict the arousal of input speech.

As the duration modeling loss $\mathcal{L}_{\text{dur}}$, we experiment with three different loss functions, all computed on the logarithm of the ground-truth durations. Let the predicted log-durations be $\widehat{\mathbf{d}} = (\widehat{d}_1, \widehat{d}_2, \ldots, \widehat{d}_U)$, and the ground-truth durations be $\mathbf{d} = (d_1, d_2, \ldots, d_U)$, where $\tilde{d}_u = \log d_u$. Specifically, the four loss functions are: (i) *mean squared error* ($\mathcal{L}_{mse}$), (ii) *mean absolute error* ($\mathcal{L}_{abs}$), and (iii) *uncertainty-based negative log-likelihood* (NLL) Loss, assuming a Gaussian distribution over log durations with predicted mean $\widehat{d}_u$ and predicted standard deviation σ_u ($\mathcal{L}_{\text{NLL}}$).

Finally, the overall speech emotion conversion loss of the architecture is as follows:

$$\mathcal{L}_{SEC} = \lambda_1 \mathcal{L}_{GAN} + \lambda_2 \mathcal{L}_{recon} + \lambda_3 \mathcal{L}_{SER} + \lambda_4 \mathcal{L}_{dur}, \quad (4)$$

where values of λ_1, λ_2 and λ_3 are adopted from [10], and λ_4 is set to 2 after a grid-search based tuning.

IV. Experimental Setup

A. Dataset

The dataset used in this study is the *in-the-wild* MSP-Podcast dataset (v1.10) [23], which contains approximately ≈238hrs of audio sourced from podcasts, with utterance-level emotion annotations provided in terms of arousal, valence, and dominance. The dataset in contrast to predominant SEC datasets (e.g., ESD [8], IEMOCAP [36]) is larger, has utterances of variable duration, has over 1400 speakers, and contains naturalistic emotional expressions. For example, the ESD contains acted-out utterances from only 10 English speakers and only ≈29 hours of acted-out utterances. To the best of our knowledge, this is one of the few works to perform SEC on an in-the-wild dataset, along with [10] and [15].

Model	DP	WVMOS ↑	SER Error ↓	
			L_{mse}	L_{abs}
HiFiGAN [10]	✗	3.26	0.084	24%
EmoConv-Diff [15]	✗	2.56	0.072	21%
MSE	✓	**3.42**	0.072	21%
L_1	✓	3.36	0.075	22%
+UnitAligner	✓	3.16	0.086	27%
Uncert	✓	3.30	**0.069**	**20**%

TABLE I: Overall performance of model versions. DP: Duration Predictor, ✓ indicates the respective model includes duration modeling, and ✗ indicates it's absence.

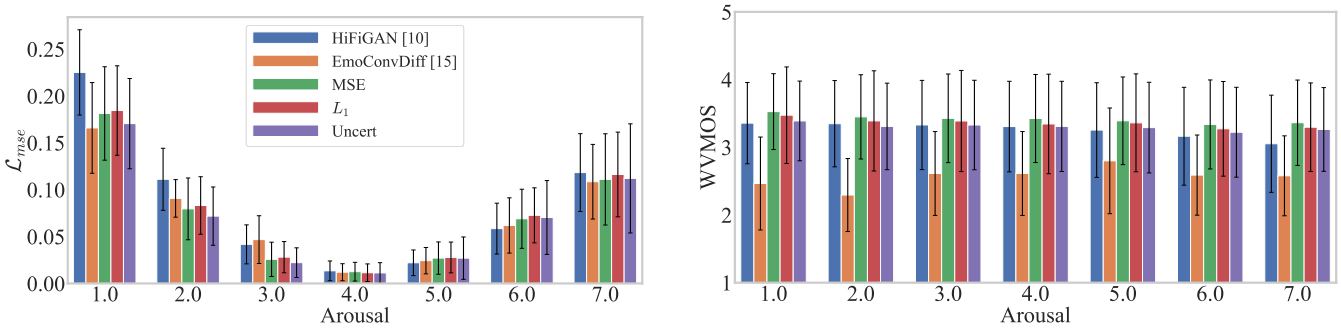
For the purpose of this work, we focus exclusively on arousal annotations for SEC, following prior works on SEC under in-the-wild conditions [3], [15]. Performing SEC on the arousal dimension, instead of categorical representation has two advantages: (i) the circumplex-model based representation better captures the subtle difference between human emotion categories [37], [38], and (ii) achieve implicit intensity control [10], as opposed to an additional effort in the categorical representation case. The arousal annotations are rated on a 1–7 scale and exhibit a distribution with a mean $\mu = 4$ and standard deviation $\sigma = 0.95$. This indicates that samples are more concentrated in the mid-range (scores 3 to 5), with fewer examples at the extremes (scores 1 and 7). This skewed distribution mirrors the nature of emotional expression in real-world, in-the-wild scenarios, such as podcasts, where extreme emotional states are relatively rare.

B. Validation Measures

We evaluate the proposed methodology based on two key aspects: its speech emotion conversion (SEC) capabilities and the naturalness of the synthesised speech. To assess SEC performance, we use mean-squared error ($\mathcal{L}_{mse}$) and mean-absolute error ($\mathcal{L}_{abs}$), both computed between the target arousal $\bar{e}$ and the SER model's prediction on the resynthesised output, $E_{SER}(\hat{y})$. For measuring the naturalness of the synthesised speech $\hat{y}$, we employ the wav2vec mean-opinion score (WVMOS) [39], an objective speech quality metric derived from wav2vec2.0 [40]. WVMOS is fine-tuned on mean-opinion scores (ranging from 1 to 5) collected through listening tests from the 2018 Voice Conversion Challenge [41], which focused specifically on naturalness. This makes WVMOS a suitable non-intrusive metric for evaluating the naturalness of $\hat{y}$. It's important to note that, since we do not use parallel data, we rely solely on non-intrusive evaluation metrics that do not require access to the ground-truth audio $\mathbf{y}_{l,s,\bar{e}}$ for emotion conversion.

C. Model Versions

As baselines for performance comparison, we use the HiFi-GAN [10] and EmoConv-Diff [15] architectures introduced earlier. Both are designed to handle in-the-wild data using the resynthesis paradigm that does not rely on parallel samples, similar to our proposed approach. Notably, neither model includes duration modeling, making them appropriate baselines



(a) SEC performance ($\mathcal{L}_{mse} \downarrow$).

(b) WVMOS $\uparrow$ performance.

Fig. 3: Speech emotion conversion performance and naturalness of generated speech across arousal levels.

for evaluating its impact. In fact, the HiFi-GAN architecture also serves as the backbone for our SEC model, into which we integrate duration modeling, further justifying its role as a baseline. We evaluate four variants of the proposed model, each employing a different approach to duration modeling,

(i) *MSE*: trains the duration predictor using mean squared error (MSE) loss.
(ii) L_1: replaces MSE with mean absolute error.
(iii) $+$*UnitAligner*: integrates the Unit Aligner module from [9], which learns discrete speech units directly from data instead of relying on pretrained HuBERT units. These learned units are then utilized by the duration predictor, improving alignment between units and acoustic frames. With the inclusion of the Unit Aligner, this baseline corresponds to a reimplementation of DurFlex [9] in our non-parallel, in-the-wild setting.
(iv) *Uncert*: introduces an uncertainty-aware duration predictor that estimates both the mean and variance of durations and is trained using Negative Log-Likelihood (NLL) loss for a probabilistic formulation.

V. Results

A. Influence of Duration Modeling

The overall performance of the different versions of the proposed model, as compared to the baselines, is shown in Table I. From the results, we observe the following: Firstly, incorporating duration modeling into the HiFiGAN baseline leads to both increased naturalness in generated speech and enhanced speech emotion conversion capabilities. The MSE variant of the duration modeling attains a WVMOS of 3.42 and a $\mathcal{L}_{abs}$ of 21%, representing an improvement over the HiFiGAN baseline, which achieves a WVMOS of 3.26 and a $\mathcal{L}_{abs}$ of 24%. Secondly, it is evident that, except for the $+$UnitAligner version, all other duration modeling approaches consistently outperform the HiFiGAN baseline, highlighting the significance of duration modeling for SEC. A probable reason why the UnitAligner does not contribute to improved duration modeling is that it is better suited for training scenarios where parallel data samples are available, as was the case in the work that originally introduced it [9], and it does not provide additional benefit in a resynthesis-based training paradigm, where direct usage of HuBERT speech units $\mathbf{z}_l$ without alignment is more appropriate. Thirdly, we note that while duration modeling yields a considerable improvement over the HiFiGAN baseline, the gains over EmoConv-Diff are relatively small. This could potentially be attributed to differences in the decoder itself, as the diffusion-based decoder used by EmoConv-Diff is more complex and has already demonstrated improvements over HiFiGAN decoders in TTS tasks [4]. Finally, among the duration modeling variants, the *MSE* and *Uncert* approaches emerge as the most effective. The *MSE* variant yields slightly better naturalness, while the *Uncert* variant achieves marginally better SEC performance. Overall, based on the empirical results, we recommend the *Uncert* variant for duration modeling due to its strong SEC performance and improved naturalness over the baseline. The SEC capabilities can be further noted in the audio examples presented online[1].

B. Performances across arousal levels

In Figure 3, the performance results are illustrated according to the target arousal level of the emotion-converted speech, considering both SEC capabilities (Fig. 3a) and the naturalness of the generated speech (Fig. 3b). Regarding SEC performance, the results in Fig. 3a indicate that incorporating duration modeling proves particularly advantageous for generating low-arousal speech, with the duration modeling variants showing a noticeably larger improvement over the baseline at low arousal levels, and only a slight improvement at high arousal. Additionally, we observe that EmoConv-Diff achieves the best SEC performance for the extreme target arousal levels of 1 and 7, with the Uncert variant of duration modeling coming closest in performance.

From Fig. 3b, we observe that duration modeling approaches consistently yield more natural-sounding speech compared to both the HiFi-GAN and EmoConv-Diff baselines. This underscores the importance of explicit duration modeling for enhancing speech naturalness. Although EmoConv-Diff demonstrates competitive SEC performance, it notably lacks

[1]https://sp-uhh.github.io/emoconv-gen

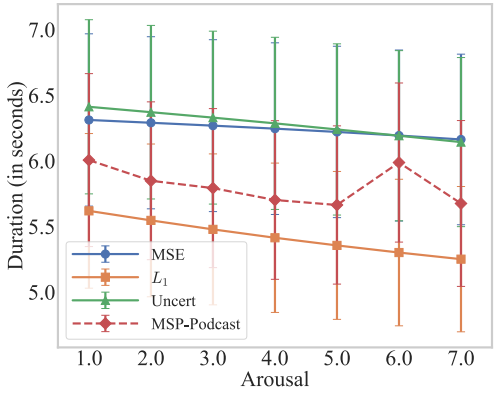


Fig. 4: Mean and standard deviation of durations for emotion-converted speech across target arousal levels. The red dashed line represents the durations from the MSP-Podcast dataset.

in naturalness. This suggests a promising direction for future work: incorporating duration modeling into diffusion-based SEC methods such as EmoConv-Diff.

C. Duration-flexible speech emotion conversion

To evaluate the effectiveness of duration modeling, Figure 4 presents the mean durations (in seconds) and their standard deviations for emotion-converted speech generated by the various model variants. Additionally, we include the oracle mean durations from the MSP-Podcast dataset (denoted by the red dashed line) corresponding to the ground-truth arousal levels. The figure clearly illustrates that all duration modeling variants successfully capture the inverse linear relationship between arousal level and speech duration: models tend to generate longer speech for low arousal levels and shorter speech for high arousal levels. This trend, also evident in the dataset reference line, is well reproduced by the SEC models incorporating duration modeling.

Among the different variants, the L_1 loss shows the most pronounced duration contrast, with the largest difference in mean duration between arousal level 1 and arousal level 7, measured as $\Delta_{1-7} = 0.37$ secs. Both the *MSE* and *Uncert* variants reflect similar patterns, with the *Uncert* variant yielding a slightly higher Δ_{1-7} of 0.21 secs. Overall, the L_1 variant tends to generate shorter duration speech compared to the other models. This behavior may stem from the nature of L_1 loss, being based on absolute error, is less sensitive to outliers (e.g., highly repetitive speech units). Consequently, it may underfit to high-repetition segments, treating them as noise and favoring shorter durations in general.

D. Modification of prosody features

To examine the prosodic modifications achieved by the duration modeling-based SEC architecture, we present Figure 5. It shows the pitch contours of the input speech (represented by black dashed lines), alongside the emotion-converted speech for target arousal level 1 (extreme low arousal, shown in blue) and arousal level 7 (extreme high arousal, shown in red). In

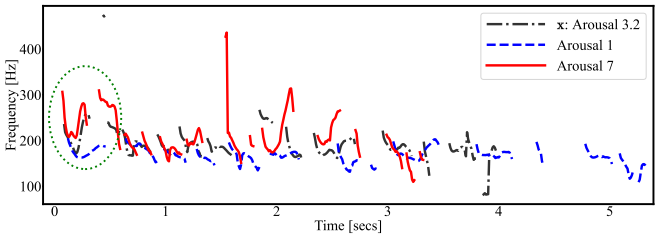


Fig. 5: Pitch contours of the input speech X, along with emotion-converted speech for target arousal levels 1 and 7.

the figure, a green dotted ellipse is used to highlight a region of interest. Due to the strong duration control demonstrated by the L_1 variant, as shown in Sec. V-C, the pitch contour analysis is conducted on speech generated by this variant.

The pitch contours in the figure reveal that the synthesised speech for high arousal ($\bar{e} = 7$) exhibits a higher mean pitch and greater pitch variability compared to both the ground-truth speech ($e = 3.20$) and the synthesised speech for low arousal ($\bar{e} = 1$). This observation is consistent with prior studies linking high emotional intensity to increased mean pitch [21], and aligns with baseline research demonstrating effective pitch control. More importantly, we observe the impact of duration modeling, which yields a shorter duration for high arousal speech (≈3.5 secs), and a longer duration for low arousal speech (≈5.3 secs), compared to the ground-truth input speech of mid-level arousal (≈ 4 secs). Finally, within the highlighted region of interest (indicated by green dotted lines), it is evident that duration modeling enables effective control and modification of speech rate. Specifically, the high arousal speech, while exhibiting a higher mean and variability in pitch, also features a noticeably shorter voiced segment (red contour) than the corresponding voiced segment in the low arousal speech (blue contour).

VI. Conclusion

In this work, we proposed a resynthesis-based duration modeling approach for speech emotion conversion that does not require parallel target speech samples—a key challenge due to the unavailability of ground-truth lexical durations during training. To overcome this, we employed a resynthesis training paradigm where the model learns to reconstruct input speech conditioned on lexical, emotion, speaker, and duration information. At inference time, emotion conversion is achieved by modifying the emotion embeddings.

We validated our approach on an in-the-wild dataset, evaluating both emotion conversion accuracy (using a pretrained SER model) and the naturalness of synthesised speech (via WVMOS). Pitch contour analysis confirms that our approach achieves not only pitch modulation but also speech rate control, producing shorter, faster speech for high arousal and longer, slower speech for low arousal. The results demonstrate the effectiveness of duration modeling, with consistent improvements in both SEC performance and naturalness over baseline methods.

REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," vol. 61, no. 5. ACM New York, NY, USA, 2018, pp. 90–99.

[2] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[3] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, 2020.

[4] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *Int. Conf. Machine Learning (ICML)*. PMLR, 2021.

[5] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-tts: A denoising diffusion model for text-to-speech," *arXiv preprint arXiv:2104.01409*, 2021.

[6] A. Triantafyllopoulos, B. W. Schuller, G. İymen, M. Sezgin, X. He, Z. Yang, P. Tzirakis, S. Liu, S. Mertes, E. André *et al.*, "An overview of affective speech synthesis and conversion in the deep learning era," *Proc. of the IEEE*, 2023.

[7] Z. Du, B. Sisman, K. Zhou, and H. Li, "Disentanglement of Emotional Style and Speaker Identity for Expressive Voice Conversion," in *Interspeech*, Sep 2022.

[8] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[9] H.-S. Oh, S.-H. Lee, D.-H. Cho, and S.-W. Lee, "Durflex-evc: Duration-flexible emotional voice conversion leveraging discrete representations without text alignment," *IEEE Tran. on Affective Computing*, 2025.

[10] N. Raj Prabhu, N. Lehmann-Willenbrock, and T. Gerkmann, "In-the-wild speech emotion conversion using disentangled self-supervised representations and neural vocoder-based resynthesis," in *Proc. ITG Conf. on Speech Comm.*, Sep. 2023.

[11] F. Kreuk, A. Polyak, J. Copet, E. Kharitonov, T. A. Nguyen, M. Rivière, W.-N. Hsu, A. Mohamed, E. Dupoux, and Y. Adi, "Textless speech emotion conversion using discrete & decomposed representations," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 2022.

[12] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Int. Conf. on Learning Representations (ICLR)*, 2014.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[14] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *Advances in Neural Inf. Proc. Systems (NeurIPS)*, 2020.

[15] N. Raj Prabhu, B. Lay, S. Welker, N. Lehmann-Willenbrock, and T. Gerkmann, "EMOCONV-Diff: Diffusion-based speech emotion conversion for non-parallel and in-the-wild data," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2024, pp. 11 651–11 655.

[16] A. N. Salman, Z. Du, S. S. Chandra, İsmail Rasim Ülgen, C. Busso, and B. Sisman, "Towards naturalistic voice conversion: Naturalvoices dataset with an automatic processing pipeline," in *Interspeech 2024*, 2024, pp. 4358–4362.

[17] F. Busquet, F. Efthymiou, and C. Hildebrand, "Voice analytics in the wild: Validity and predictive accuracy of common audio-recording devices," *Behavior Research Methods*, 2023.

[18] K. Zhou, "Emotion modelling for speech generation," PhD thesis, National University of Singapore, 2022, available at https://scholarbank.nus.edu.sg/handle/10635/243782.

[19] C. Robinson, N. Obin, and A. Roebel, "Sequence-to-sequence modelling of f0 for speech emotion conversion," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2019, pp. 6830–6834.

[20] T.-H. Kim, S. Cho, S. Choi, S. Park, and S.-Y. Lee, "Emotional voice conversion using multitask learning with text-to-speech," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2020, pp. 7774–7778.

[21] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Tran. on Affective Computing*, 2023.

[22] K. Zhou, B. Sisman, and H. Li, "Vaw-gan for disentanglement and recomposition of emotional elements in speech," in *IEEE Spoken Language Tech. Workshop*, 2021.

[23] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Tran. on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[24] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 32, 2019.

[25] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis." in *Interspeech*, 2017, pp. 4006–4010.

[26] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2018, pp. 4779–4783.

[27] Z. Zeng, J. Wang, N. Cheng, T. Xia, and J. Xiao, "Aligntts: Efficient feed-forward text-to-speech system without explicit alignment," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*. IEEE, 2020, pp. 6714–6718.

[28] W.-C. Huang, Y.-C. Wu, and T. Hayashi, "Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2021, pp. 5944–5948.

[29] S. Wang, T. Qi, C. Lu, Z. Luo, and W. Zheng, "Enhancing zero-shot emotional voice conversion via speaker adaptation and duration prediction," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2025.

[30] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Interspeech*, 2021.

[31] W.-N. Hsu, T. Remez, B. Shi, J. Donley, and Y. Adi, "Revise: Self-supervised speech resynthesis with visual input for universal and generalized speech regeneration," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023.

[32] D. de Oliveira, N. Raj Prabhu, and T. Gerkmann, "Leveraging semantic information for efficient self-supervised emotion recognition with audio-textual distilled models," in *Interspeech*, 2023.

[33] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.

[34] L. I.-K. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, vol. 45, no. 1, p. 255, Mar. 1989.

[35] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Tran. on Pattern Analysis and Machine Int.*, 2023.

[36] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, 2008.

[37] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, 1980.

[38] N. Raj Prabhu, N. Lehmann-Willenbrock, and T. Gerkmann, "End-to-end label uncertainty modeling in speech emotion recognition using bayesian neural networks and label distribution learning," *IEEE Tran. on Affective Computing*, vol. 15, no. 2, pp. 579–592, 2024.

[39] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: A unified framework for bandwidth extension and speech enhancement," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2023, pp. 1–5.

[40] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, pp. 12 449–12 460, 2020.

[41] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z.-H. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Interspeech*, Apr 2018.

5

Discussion and Conclusions

5.1 Main Contributions of this Thesis

In this chapter, I summarize and reflect on the main contributions of this thesis in light of its overarching goal: advancing a social signal processing approach to enable social intelligence in intelligent agents. The discussion is organized according to the research questions outlined in Section 1.7. First, under the first pillar of social signal processing—*recognition of affective expressions*—I discuss contributions related to probabilistic approaches for modeling label uncertainty in individual-level affect recognition (Thesis Contribution 1). This particular contribution directly addresses research questions **RQ1** and **RQ2**. Second, still within the pillar of *recognition of affective expressions*, I analyze contributions concerning the modeling of affect at the group level (Thesis Contribution 2). This contribution corresponds to the research questions **RQ3** and **RQ4**. Finally, turning to the second pillar of social signal processing—*synthesis of affective expressions*—I evaluate contributions related to generative modeling for emotion-conditioned speech synthesis (Thesis Contribution 3). This particular contribution addresses the research questions **RQ5**, **RQ6**, and **RQ7**.

5.1.1 Recognition of Affective Expressions

Probabilistic Approach to Label Uncertainty Modeling

Modeling label uncertainty in affect recognition arises from the challenge of defining a reliable "ground truth." Affect annotations are inherently subjective and ambiguous, so collapsing multiple annotator perspectives into a single consensus—through averaging, majority voting, or annotator training—loses valuable information about human disagreement. Thesis Contribution 1 focuses on systematically capturing this uncertainty in individual-level affect recognition. In [P1], we introduced a probabilistic framework combining Bayesian Neural Networks (BNNs) with Label Distribution Learning (LDL). By training directly on annotation distributions using a LDL-based KL divergence loss, our approach captures both the central tendency and variability in affect annotations. Under a Gaussian assumption, the predictive mean and standard deviation are fine-tuned via the KL loss, and empirical results confirm that this probabilistic approach improves uncertainty estimation while maintaining competitive mean prediction. Beyond technical performance, this work reframes "ground truth" as a distribution of possible interpretations rather than a single, objective label [P6].

However, modeling uncertainty introduces a trade-off: improved uncertainty estimates often come with a modest reduction in average emotion prediction accuracy. This limitation is

accentuated in datasets with sparse and skewed annotations, such as RECOLA [156], MSP-Podcast [42], and MSP-Conversation [115], where small numbers of annotators (3–6) and diverse contexts lead to asymmetric, outlier-prone distributions. To better capture these characteristics, we proposed modeling annotations with a *Student's t-distribution*, which naturally accommodates heavy tails and adjusts for the number of available annotations. This formulation yields a novel mean-seeking KL loss between the t-distribution ground truth and the model's Gaussian predictions, improving robustness in low-sample settings.

Empirical evaluation demonstrates that the t-distribution–based loss mitigates the trade-off observed under Gaussian assumptions: it simultaneously enhances label uncertainty estimation and mean prediction, accelerates convergence, and reduces overfitting ([P1]; Table 1, Figures 6–8). Benefits are particularly evident in complex datasets and cross-corpus experiments, and the approach performs best when inter-annotator correlation is high but annotations are limited. Collectively, these findings highlight that explicitly modeling label uncertainty—accounting for subjectivity, ambiguity, sparsity, and outliers—provides a more accurate, robust, and psychologically plausible framework for individual-level affect recognition.

Implications of Thesis Contribution 1 The findings from Thesis Contribution 1 have several important implications for affective computing and the broader study of SSP. First, by explicitly modeling label uncertainty, this work challenges the conventional notion of "ground truth" in emotion recognition. Most prior research collapses annotator disagreement into a single value, implicitly treating subjectivity and ambiguity as noise. Our probabilistic and distribution-based approach demonstrates that these sources of variation contain meaningful information that can improve predictive performance, robustness, and interpretability.

Second, the integration of BNNs with LDL highlights a practical strategy for incorporating uncertainty into affect modeling. This addresses a gap in the literature where previous methods either modeled uncertainty deterministically or assumed Gaussian distributions that poorly reflect the sparsity, skewness, and occasional outliers typical of real-world annotations. By introducing the Student's t-distribution to accommodate such annotation characteristics, our work provides a principled framework that more faithfully represents human affective perception, particularly in settings with limited or unevenly distributed annotations.

Third, these findings underscore a broader conceptual shift: affect should be understood as inherently probabilistic and context-dependent, rather than fixed and absolute. This has implications for the design of socially intelligent systems, suggesting that robust perception requires models that capture both the central tendencies and the variability of human judgments. From a methodological perspective, this contribution provides a template for future research to explicitly incorporate annotation uncertainty, offering a bridge between computational modeling and psychological theory. Taken together, Thesis Contribution 1 establishes that modeling label uncertainty is not only technically advantageous but also conceptually important: it reframes affect recognition as a richer, ethically grounded, and more psychologically plausible task.

Research Questions

RQ1 *How can we best model label uncertainty for individual-level emotion recognition, and what empirical gains—beyond improved characterization of uncertainty in emotion annotations—does this modeling yield?*

Label uncertainty for individual-level emotion recognition can be effectively modeled by combining Bayesian Neural Networks (BNNs) with Label Distribution Learning (LDL). In this formulation, the BNN produces stochastic outputs that approximate a predictive distribution, while LDL leverages a KL-divergence loss to train directly on the annotation distributions rather than collapsing them into point estimates. This integration allows the model to capture both central tendency and variability in annotations, reflecting the multiplicity of human interpretations. Empirically, the combined BNN+LDL approach yields clear, statistically significant gains over deterministic baselines such as single-task learning (STL) and MTL: it provides more accurate uncertainty estimation, improves mean prediction, and enhances robustness to annotator disagreement. In addition, ablation studies show that BNNs and LDL contribute complementary strengths, with the combination outperforming either method alone. Beyond technical improvements, this approach reframes the notion of "ground truth" as a distribution of possible interpretations, offering a richer and more interpretable basis for modeling affect.

RQ2 *How can label uncertainty be modeled to robustly account for sparse and limited annotations, beyond the common assumption of Gaussian-distributed annotations?*

Limited, sparse, and skewed annotations can be more robustly modeled with a *Student's* t-distribution, which combines heavier tails to accommodate outliers with degrees of freedom that encode the number of available annotations. This provides a principled way to tie annotation availability directly to label uncertainty, overcoming the limitations of Gaussian assumptions in affect datasets that typically include only 3–6 annotators per sample. In our formulation, the Gaussian predictive distribution of a Bayesian neural network is trained against the t-distribution ground truth using a mean-seeking KL-divergence objective, ensuring both tractability and full-distribution modeling. Empirical findings show that this approach improves uncertainty estimation and mean prediction simultaneously, mitigates the trade-off observed in Gaussian models, accelerates convergence, and yields stronger robustness in complex datasets and cross-corpus evaluations.

Having examined uncertainty modeling in individual-level affect recognition, the next step in this thesis is to extend the scope from individuals to groups. This transition is motivated by the fact that social interactions rarely unfold at the level of isolated individuals; instead, affect emerges through interpersonal dynamics, coordination, and shared context. Consequently, recognizing affect at the group level introduces additional challenges—such as modeling relational structure, temporal dependencies, and collective processes—that go beyond individual-level modeling alone. These considerations motivate the following section, which focuses on graph-based multimodal modeling for group-level affect.

Graph-based Multimodal Modeling of Group-level Affect

Thesis Contribution 2 addresses the recognition of affect at the group level, which presents challenges beyond individual-level modeling due to its dynamic, relational, and emergent nature. Existing GER research often simplifies group affect by treating it as a static aggregate of individual states or by focusing on non-purposive groups (e.g., crowds, protests), leading to a persistent *theory–method misalignment* between computational approaches and organizational psychology.

To overcome these limitations, we make two key contributions: (i) we introduce novel group affect annotations on purposive group interactions, guided by theory-driven strategies from organizational psychology; and (ii) we propose a graph-based multimodal modeling framework that captures group affect dynamics, emphasizing processes of convergence and divergence among members [P4]. The annotation protocol aligns window lengths with theoretically relevant constructs, recruits trained annotators with domain expertise, and captures the ebb and flow of affect over time, producing reliable, ecologically valid data for computational modeling.

The graph-based framework represents group interactions as graphs, with nodes corresponding to participants and edges encoding interpersonal relationships. Temporal dependencies are modeled via an LSTM applied to graph features, while attention-based learning estimates edge weights to capture relational dynamics. This design allows the model to scale to arbitrary group sizes and reflect both convergence and divergence in affective states. Experiments show that this approach outperforms handcrafted baselines and that multimodal integration (audio–visual) provides the strongest performance. Analyses further reveal systematic patterns: convergence among members is associated with stronger positive or negative affect, whereas divergence corresponds to neutral affect [P4].

Research Questions

RQ3 *In what ways can group-level affect annotations be advanced beyond existing approaches to achieve stronger alignment between theories in organizational psychology and methods in affect recognition?*

Advancing group-level affect annotations requires explicitly addressing the persistent *theory–method misalignment* between affect recognition research and organizational psychology. Existing datasets often reduce group affect to static snapshots or aggregated individual states, thereby neglecting its dynamic, collective, and emergent nature. To move beyond these limitations, we designed an annotation strategy grounded in organizational psychology and tailored for purposive group interactions. This strategy involved (i) iteratively tuning annotation windows to align with constructs central to group affect dynamics, and (ii) recruiting and training annotators from organizational psychology programs to ensure conceptual expertise and reliability. The resulting annotations capture the *ebb and flow* of affect across interactions, reflecting processes of convergence and divergence among group members. By integrating theoretical insights from psychology with methodological advances in affect recognition, this approach produces annotations that are both more faithful to the social construct and better suited for computational modeling, thereby advancing interdisciplinary alignment in the study of group affect.

Research Questions

RQ4 *To what extent can group-level constructs, such as collective affect, be modeled in a data-driven framework that accommodates arbitrary group sizes and captures the inherent dynamics of affect?*

Group-level constructs such as collective affect can be effectively modeled in a data-driven framework by adopting graph-based architectures inspired by social network theory. Unlike traditional approaches that depend on handcrafted features (e.g., synchrony or mimicry), our framework meets two essential requirements: (i) scalability to arbitrary group sizes, achieved by representing members as graph nodes and their relationships as edges, and (ii) the ability to capture dynamic processes of convergence and divergence through temporal modeling with an LSTM layer applied to graph-derived features. Experiments demonstrate that this approach consistently outperforms handcrafted baselines, while multimodal integration of audio and visual signals further strengthens predictive performance. Analyses reveal systematic patterns: convergence among group members is associated with stronger affective states (positive or negative), whereas divergence tends toward neutral affect. Taken together, these results show that group-level constructs can be robustly captured through a principled, data-driven, and multimodal modeling framework that accommodates both the scale and dynamics of real-world group interactions.

Implications of Thesis Contribution 2 The findings from Thesis Contribution 2 have several important implications for affective computing and the study of SSP at the group level. First, by extending affect modeling from individuals to groups, this work demonstrates that group affect is best understood as an emergent, relational property rather than a static aggregation of individual states. Whereas prior GER research often oversimplified group affect, our graph-based, multimodal framework captures the dynamic interplay of convergence and divergence among group members, highlighting the temporal and relational dependencies that shape collective emotional states. This addresses a long-standing gap in the literature and provides a computational model that aligns more closely with psychological theory.

Second, the introduction of theoretically grounded annotations illustrates the importance of interdisciplinary alignment between organizational psychology and affective computing. By designing annotation protocols tailored for purposive group interactions, recruiting trained annotators with domain expertise, and iteratively tuning annotation windows to reflect relevant constructs, our work establishes a standard for producing reliable, ecologically valid group affect data. This methodological rigor allows computational models to capture the ebb and flow of group emotions, supporting more robust and interpretable predictions and providing a bridge between theory and practice.

Third, the use of graph-based architectures demonstrates a practical strategy for modeling multilevel, multimodal affective phenomena. Representing participants as nodes and their interactions as edges allows the framework to scale to arbitrary group sizes, while attention-based learning and temporal modeling capture convergence and divergence in real time. Empirical results show that this approach outperforms traditional, handcrafted feature-based models and reveals systematic patterns of affective dynamics, emphasizing the value of data-driven relational structures for understanding social interactions.

Together, the Thesis Contribution 2 advances the *perception* dimension of the overarching aim of this thesis by extending affect modeling from individuals to groups and, crucially, by showing why such an extension requires both theoretical and methodological alignment. Taken in combination with the previous section on uncertainty modeling, our findings highlight that affect in naturalistic settings is inherently multilevel, relational, and uncertain. At the individual level, uncertainty arises from subjective interpretation; at the group level, it additionally emerges through interpersonal dynamics, convergence, and divergence. - which more clear and puts thesis contribution 1 and 2 together under the perception aspect. More broadly, these findings situate group affect modeling as a critical component of socially intelligent systems. Perceiving affect at multiple levels—while accounting for uncertainty and interpersonal dependencies—provides a more ecologically valid foundation for downstream applications such as team-support systems, collaborative AI, and affect-aware agents.

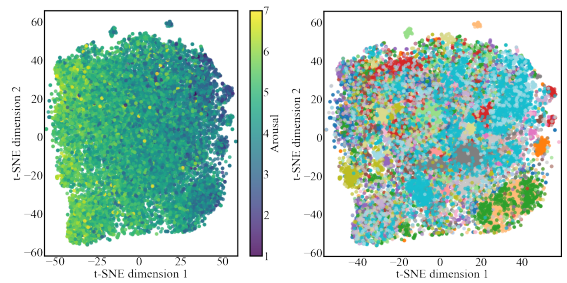
While the preceding sections focused on recognizing affect at both the individual and group levels, affective computing and SSP encompasses *not only the ability to perceive* but also to *synthesise* affect. Understanding how affect is expressed allows machines to interpret human behavior, whereas generating affective signals enables socially aligned communication. Having addressed perception-oriented challenges, including label uncertainty and relational dynamics, the thesis now turns to the second pillar: the synthesis of affective expressions through generative modeling.

5.1.2 Synthesis of Affective Expressions

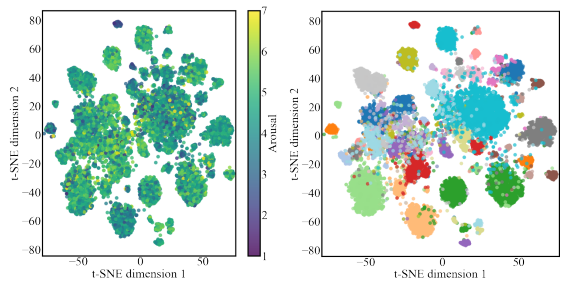
Generative Models for Emotional Speech Synthesis with In-the-wild Data

For over two decades, emotion-conditioned speech synthesis (ESS) has been a sustained area of research, positioned as a closely related yet distinct subfield within speech synthesis and text-to-speech (TTS), with strong ties to affective computing. A key limitation, however, is that most data-driven methodologies have been developed and evaluated using acted emotional speech data, which lack ecological validity. Unlike acted data—essentially read speech—*in-the-wild* recordings are spontaneous, capturing diverse speaking styles, emotional expressions, nonverbal cues (e.g., laughter, lip smacks), and disfluencies such as hesitations or interruptions [42], [57]. The importance of shifting toward in-the-wild datasets has been consistently emphasized as a limitation of current research and a critical direction for future work in doctoral dissertations [56] and literature reviews [55], [182]. To the best of our knowledge, our contributions [P9]–[P11] are the first to explicitly tackle ESS in the context of in-the-wild datasets.

Two central research challenges were identified early on: (i) the absence of parallel speech data in in-the-wild settings, which complicates both training and evaluation, and (ii) the high variability and noise in naturalistic recordings—including uncontrolled acoustic conditions, speaker diversity, and emotion variability—which make reliable modeling particularly challenging. To address the lack of parallel data, we proposed a SSL-based disentanglement and resynthesis framework that forms the backbone of our subsequent contributions. Here, speech is decomposed into task-specific embeddings: speaker embeddings $\mathbf{z}_s$ from speaker verification (SV) models, lexical embeddings $\mathbf{z}_l$ from ASR, and emotion embeddings $\mathbf{z}_a$ from SER. These disentangled representations serve as inputs for reconstructing the waveform, and at inference time, the emotion embedding can be manipulated to generate speech conditioned on a target emotion. To assess whether pretrained speech representations provide a suitable basis for disentangling affective and speaker-related information, we used t-distributed Stochastic



(a) Color-coded for Emotion-level (Left) and Speaker ID (Right)



(b) Color-coded for Emotion-level (Left) and Speaker ID (Right)

Figure 5.1: t-SNE plots for the Emotion Embedding, $\mathbf{z}_a$ (a) and Speaker Embedding, $\mathbf{z}_s$ (b).

Neighbor Embedding (t-SNE) visualizations to qualitatively inspect how emotion and speaker attributes are organized in the embedding space. As shown in Figure 5.1a, the emotion embeddings capture meaningful affective variation, with emotion levels gradually decreasing from left to right in the left plot. In contrast, the corresponding speaker IDs appear largely scattered, with only a few clusters emerging for speakers who *consistently express similar emotion* levels. A complementary pattern is visible in the speaker embeddings (Figure 5.1b): clear clustering is observed by speaker ID in the right plot, whereas emotion levels are more diffusely distributed. Rather than indicating strict class separability, these patterns show that the pretrained encoders already encode partially disentangled structure—emotion cues align along smoother gradients, while speaker identity forms more distinct clusters. This supports the use of pretrained models as a foundation for downstream disentanglement in our synthesis framework. At the same time, the diffuse and overlapping clusters highlight the complexity of in-the-wild data, where emotional and speaker-related signals are naturally less cleanly separated than in acted corpora. This motivates the need for modeling approaches capable of handling such noisy, organically occurring affective structure.

Implications of Thesis Contribution 3 Thesis Contribution 3 advances the *synthesis* dimension of the overarching aim of this thesis by demonstrating how emotion-conditioned speech synthesis can be achieved from naturalisitc, in-the-wild data. By leveraging disentangled representations of speaker, lexical, and emotion content, the framework leverages generative modeling to enable principled control over affective expression without requiring parallel corpora—a key limitation in prior ESS research.

The progression from HiFiGAN-based resynthesis [P9] to diffusion-based modeling [P10] shows that generative approaches can robustly handle high variability and rare emotion instances inherent in naturalistic recordings. Diffusion models enhance robustness, reduce mode collapse, and improve coverage of extreme emotional states, addressing limitations of conventional vocoder-based synthesis. Incorporating duration modeling [P11] further captures prosodic features such as rhythm and stress, enabling speech generation that aptly reflects human affective timing and dynamics. The findings from Thesis Contribution 3 establish that affective speech synthesis is not only technically feasible under uncontrolled, naturalistic conditions but also conceptually significant: it provides an ecologically valid, methodologically grounded pathway to generate emotionally expressive speech. This has direct implications for socially intelligent systems, allowing synthesized speech to be affectively appropriate, contextually aligned, and perceptually natural in real-world interactions.

By integrating the synthesis aspect of SSP, this contribution demonstrates how affective

computing can move beyond recognition toward expressive, socially grounded generation. Through disentangled embeddings, diffusion-based models, and duration control, the framework produces speech that captures natural variability in prosody, pitch, and rhythm. This enables applications such as a collaborative robot participating in purposive group tasks, where synthesized speech communicates affective states aligned with the interaction context. Overall, Thesis Contribution 3 provides both a methodological and conceptual foundation for machines to generate perceptually coherent and socially meaningful affective behaviors.

Research Questions

RQ5 *How can generative modeling approaches be applied to the task of speech emotion conversion in the challenging context of in-the-wild affect datasets?*

Generative modeling for speech emotion conversion (SEC) on in-the-wild datasets can be realized via a disentanglement–resynthesis framework that does not require parallel data. We extract task-specific embeddings—speaker ($\mathbf{z}_s$), lexical ($\mathbf{z}_l$), and emotion ($\mathbf{z}_a$)—using pretrained SSL models (SV, ASR, SER), and train a decoder to reconstruct waveforms from these compact representations. By manipulating $\mathbf{z}_a$ at inference, the system generates speech conditioned on a target emotion while preserving speaker identity and linguistic content. A modified HiFiGAN vocoder serves as the decoder, operating directly on disentangled embeddings with adversarial training. Empirically, progressively adding disentangled factors improves quality and naturalness, larger context windows benefit emotion control, and pitch analyses confirm successful arousal conditioning. This establishes a practical, ecologically valid path to emotion conversion on in-the-wild corpora.

RQ6 *To what extent can diffusion models, given their probabilistic and generative modeling advantages over neural vocoders, enhance the capabilities of emotion-conditioned speech synthesis?*

Diffusion models enhance emotion-conditioned speech synthesis by explicitly modeling the full conditional data distribution and mitigating mode collapse commonly observed in neural vocoders. In our work, replacing the HiFiGAN decoder with a diffusion-based decoder (*EmoConv-Diff*) within the disentanglement–resynthesis pipeline enabled broader coverage of rare and extreme emotion regions, alongside improved prosodic expressivity and stronger pitch control. Operating on phoneme-level, speaker- and affect-independent Mel features further aligned with diffusion inductive biases, stabilizing training and reducing reliance on parallel data. However, these gains come with a trade-off: while diffusion models yield superior emotion rendering—particularly at arousal extremes—they tend to produce speech that is perceptually less natural than vocoder-based baselines. This highlights both their potential for controllability and robustness, and the need for future work to balance expressivity with naturalness.

RQ7 *How does the integration of a dedicated duration modeling module affect the performance of speech emotion conversion, and can such a module be trained on in-the-wild data without relying on ground-truth duration information from parallel speech corpora?*

Research Questions

Integrating a dedicated duration module substantially improves speech emotion conversion by enabling control over rhythm and stress—prosodic dimensions that strongly influence perceived emotion. Using a lightweight convolutional neural network (CNN)–linear predictor trained on HuBERT-based discrete units, duration is modeled as token repetitions ($\mathbf{d}_i$) with a simple MSE loss, eliminating the need for parallel corpora or explicit ground-truth durations. This allows the system to generate shorter, faster speech for high arousal and longer, slower speech for low arousal, as validated through pitch and duration analyses. Crucially, adding this module not only enhances naturalness and controllability but also boosts the SEC performance of the HiFiGAN-based pipeline to match, and in some cases surpass, that of the diffusion-based decoder. These findings demonstrate that duration can be effectively learned from in-the-wild data and is a critical missing component for faithful and expressive emotional speech synthesis.

Summary of Overall Contribution Together, the three contributions of this thesis form a coherent framework for socially intelligent computational systems, advancing both perception and synthesis of affective signals across multiple levels of analysis.

- **Thesis Contribution 1:** Probabilistic modeling of label uncertainty at the individual level addresses the inherent subjectivity and ambiguity of affect annotations. By reframing disagreement not as noise but as informative variation, this work provides more robust, interpretable, and psychologically grounded models for individual affect recognition.
- **Thesis Contribution 2:** Graph-based multimodal modeling of group affect extends perception from individuals to collective settings. By aligning annotation protocols with organizational psychology theory and capturing convergence/divergence dynamics through data-driven relational modeling, this contribution demonstrates that group affect is emergent and relational, rather than a static aggregation of individual states.
- **Thesis Contribution 3:** Generative modeling of emotional speech with in-the-wild data addresses the synthesis aspect of affective computing. Through disentangled representations, diffusion-based generation, and duration modeling, this work enables the production of contextually appropriate, ecologically valid, and expressive speech, laying the groundwork for socially interactive agents capable of generating nuanced affective behaviors.

Collectively, these contributions advance the methodological, conceptual, and practical foundations of affective computing. Methodologically, the thesis introduces probabilistic, relational, and generative frameworks that explicitly account for subjectivity, interpersonal relationships, and naturalistic variation in affective expressions. Conceptually, it frames affect as a dynamic, socially embedded, and context-dependent phenomenon, while respecting its temporal dynamics, multidimensionality, and multilevel organization, and maintaining the ecological validity of affective expressions. Practically, the work demonstrates how computational models can move beyond recognition to generation, enabling machines to both perceive and express affect in ways that are socially aligned and contextually meaningful. By integrating perception and synthesis across multiple levels—from individual affective states to group-level dynamics—this thesis establishes a comprehensive framework for social signal processing, laying the groundwork for adaptive, socially intelligent systems capable of understanding and generating affect

in real-world interactions.

5.2 Directions for Future Research

By advancing the study of affect across recognition and synthesis, this thesis opens important questions that point toward several avenues for future research.

Label Uncertainty in Group-level Affect

In this thesis, and specifically in [P1], label uncertainty was studied in the context of *individual-level* affect annotations $\{\mathbf{a}_t^{(i)}\}_{i=1}^K$. However, this was not extended to our subsequent work on *group-level* affect recognition. More broadly, the question of how to incorporate annotation uncertainty is not unique to group-level affect; it is equally relevant for other constructs involving ambiguous or socially interpreted labels, such as engagement, rapport, or conversational dynamics. Thus, while this section is framed around group affect, the methodological considerations extend naturally to a wider class of social signal processing tasks.

In [P4], the ground truth of group affect was derived using the EWE (1.14) and the proposed graph-based model relied on a deterministic mapping between interlocutor cues $\{\mathbf{X}_{t_1:t_2}^{(i)}\}_{i=1}^M$ and $\mathbf{a}_t^{(\text{EWE})}$ (see Equations (1.34) and (1.35)). This framework, through graph parameters effectively captured both the spatial (interpersonal relationships) and temporal (interaction dynamics) dimensions of group affect, but it *remained uncertainty-agnostic*. Our focus at that stage was on introducing temporal modeling into group affect, leaving uncertainty-aware extensions as a natural next step.

A straightforward extension would be to replace the deterministic regressor $f_{\mathbf{w}}$ in (1.35) with a BBB-based probabilistic regressor, yielding stochastic outputs:

$$\mathbf{a}_t^{(g)} \sim P(\cdot \mid \mathcal{G}_\psi; \mathbf{w}). \tag{5.1}$$

While this introduces label uncertainty, it does so only at the regressor level, without propagating uncertainty through the graph encoder $\mathcal{G}_\psi$ itself. *This limitation motivates the need for a more holistic treatment of uncertainty: if uncertainty exists not just in the labels but also in how nodes, edges, and the overall graph structure are represented, then modeling only the output uncertainty provides an incomplete picture.* A recent review of uncertainty-aware GNNs [290] highlights that predictive performance and robustness can be enhanced by modeling uncertainty at three complementary levels: (1) node-level, (2) edge-level, and (3) graph-level.

At the **node-level**, uncertainty in *individual cues* can be modeled by treating each node embedding as a distribution rather than a point estimate. For example, [291] combine a GNN with a variational graph auto-encoders (VGAE) [292] to represent each node feature matrix as a distribution in the latent space, $\mathcal{N}(\mu_i, \sigma_i^2)$, rather than a deterministic vector, $\mathbf{X}_i$. This variational treatment enhances the model's ability to quantify the uncertainty associated with node-level embeddings.

At the **edge-level**, uncertainty in *interpersonal relationships* can be captured by placing distributions over edge weights, for example by extending the learnable attention weights in [P4] (Equation 1.37) into a probabilistic form: $\alpha_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$, where α_{ij} represents the relationship strength between interlocutors i and j, with learned mean μ_{ij} and variance σ_{ij}^2. For instance, [293] proposed *DropEdge*, which randomly removes edges by sampling Bernoulli

random variables during each training iteration, essentially similar to MC *dropout* based uncertainty modeling [202].

At the **graph-level**, the aggregated representation can itself be treated as a realization from a parametric random graph family by integrating Bayesian inference directly into the GNN. Following [294], the observed graph is modeled as a sample $\mathcal{G}$ from a random graph distribution parameterized by ψ, and inference is performed jointly over the graph parameters and node labels:

$$P(\mathcal{G}, \mathbf{a}^{(g)} \mid \{\mathbf{X}^{(i)}\}_{i=1}^{M}, \psi). \tag{5.2}$$

This formulation enables uncertainty-aware predictions at the level of the entire group structure by treating both the graph topology and the group-level affect label as random variables. Such an approach allows the model to jointly capture uncertainty in how the group is structured and how its collective affect emerges.

Together, node-, edge-, and graph-level formulations provide multiple pathways for advancing group affect recognition by embedding uncertainty-awareness directly into graph-based modeling.

Latent Speech Representations-based Diffusion Models for SEC

Within the task formulation of SEC in Equations (1.44) and (1.45) (Section 1.6), the lexical representation $\mathbf{z}_l$—which encodes the temporal evolution of linguistic content—forms the backbone on which decoders such as HiFiGAN (Section 4.1 [P9]) and diffusion models (Section 4.2 [P10]) operate. In our HiFiGAN experiments, $\mathbf{z}_l$ was represented as discrete HuBERT speech units, whereas in EMOCONV-Diff it was replaced by "average-voice" phoneme-level Mel features to align with prior diffusion-based speech synthesis work [64]. However, this design choice led to practical challenges. Specifically, as shown in [P11], duration modeling proved far easier with discrete HuBERT units—where duration reduces to predicting repetitions of tokens—than with phoneme-level Mel features, which exhibit inherent time-dependent variation. Attempts to jointly train a duration predictor with EMOCONV-Diff produced unstable training and unintelligible outputs. Consequently, duration modeling was only introduced with HiFiGAN in [P11], where it improved SEC performance beyond EMOCONV-Diff. In the same work [P11], we showed that, without duration modeling in HiFiGAN, EMOCONV-Diff showed superior SEC capabilities over HiFiGAN, but it did so at the cost of reduced naturalness of generated speech. These results suggest that extending duration modeling to EMOCONV-Diff could mitigate its naturalness drawbacks, but doing so requires revisiting the design of $\mathbf{z}_l$.

Latent Diffusion Models. To overcome the limitations of phoneme-level Mel features, future work could adopt a latent diffusion modeling strategy, reformulating the generation process as:

$$\begin{aligned} &\mathbf{X}_{\text{trg}} \sim P_{\mathbf{w}}(\mathbf{X} \mid \mathbf{z}_l^{\text{latent}}, \mathbf{z}_s, \mathbf{z}_a, \mathbf{a}_{\text{trg}}), \\ &\text{where } \mathbf{z}_l^{\text{latent}} = E_\theta(\mathbf{X}) \end{aligned}$$

where $\mathbf{z}_l^{\text{latent}}$ is a compact latent representation of the input speech obtained via a learned encoder E_θ, instead of raw phoneme-level Mel features. By modeling diffusion in this compressed latent space rather than in the high-dimensional acoustic space, training can become more stable, duration modeling more tractable, and naturalness of synthesis potentially restored.

Building on this formulation, several promising avenues emerge for future research. A central challenge lies in balancing discrete and continuous latent representations ($\mathbf{z}_l^{\text{latent}}$) to jointly enable stable duration modeling and expressive emotion conditioning. Another important direction is to evaluate the extent to which latent diffusion models can mitigate the trade-off between naturalness and emotional expressivity observed in HiFiGAN and EMOCONV-Diff. Furthermore, probabilistic formulations in the latent space may offer a principled way to capture uncertainty in emotional prosody, particularly under in-the-wild conditions where annotations are sparse or noisy. Collectively, these directions highlight latent diffusion modeling as a promising pathway for advancing ESS and SEC toward more natural, controllable, and robust affective speech generation.

Joint control of arousal and valence for SEC

With respect to the multilevel nature of affect, this thesis adopts the circumplex model to quantify affect, which maps emotions onto a two-dimensional orthogonal plane defined by arousal and valence (Figure 1.2). In this representation, the two dimensions jointly capture affective states: for example, "happy" and "sad" may share moderate arousal but differ in valence, while "calm" and "angry" may both be low in arousal but opposite in valence. However, a limitation of the contributions in [P9]–[P11] is that controlled synthesis focused only on the arousal dimension, without addressing valence.

Future work should extend emotion-conditioned synthesis to both dimensions. A crucial challenge here is the empirical correlation between arousal and valence in affect annotations. Of note, in [P9]–[P11], initial experiments reveal that generating speech of high-arousal also resulted in increasing the valence to an extent, both perceptually and in evaluations based on a SER model. While Russell's circumplex model [104] posits orthogonality, in practice, annotated datasets show strong correlations because certain emotional states are more frequently expressed. For example, high-arousal positive emotions (e.g., excitement) and low-arousal negative emotions (e.g., sadness) are far more commonly encountered than high-arousal negative emotions (e.g., anger) or low-arousal positive emotions (e.g., calmness). This results in class imbalance, where annotated datasets such as RECOLA [156] and MSP-Podcast [42] are skewed toward correlated states of arousal and valence. Consequently, encoders trained on these corpora tend to embed this correlation directly into the learned emotion representations, and loss functions optimized on these embeddings reinforce the imbalance by penalizing less frequent, decorrelated states.

Methodologically, future research could address this by incorporating disentangled latent representations that explicitly separate arousal and valence dimensions, while regularizing against correlation in the latent space. Variational frameworks with structured priors, such as β-VAEs [295] or FactorVAEs [296], offer principled ways to encourage independence between dimensions, while recent work on correlation-aware contrastive objectives [297] could be adapted to penalize over-reliance on dataset biases. In the context of generative modeling, diffusion models conditioned on disentangled embeddings may provide a pathway to synthesize less frequent states such as high-arousal/low-valence or low-arousal/high-valence emotions more reliably, thereby reducing bias and improving ecological validity in emotion-conditioned speech synthesis.

Unified Affect-understanding and Synthesis Loop for Interactive Agents

A promising direction for future research lies in bringing together affect recognition and affect synthesis within a unified computational framework. In this thesis, these two pillars

were deliberately explored in depth but separately—label uncertainty modeling for affect recognition, and generative modeling for affective speech synthesis. Integrating them, however, opens the door to a richer class of socially intelligent systems. Recognition and synthesis are inherently *interdependent*: effective synthesis of affective behavior benefits from an accurate understanding of the affective context, while recognition models can be strengthened through feedback signals originating from generative processes. Studying them jointly would therefore allow for a more holistic treatment of how interactive agents perceive, model, and express affect. Future research may move toward closed-loop systems that seamlessly combine these two capabilities, enabling agents that both interpret and express affect in socially aligned, adaptive, and interactive ways. An illustration of such a unified affect-understanding and synthesis loop is provided in Figure 5.2.

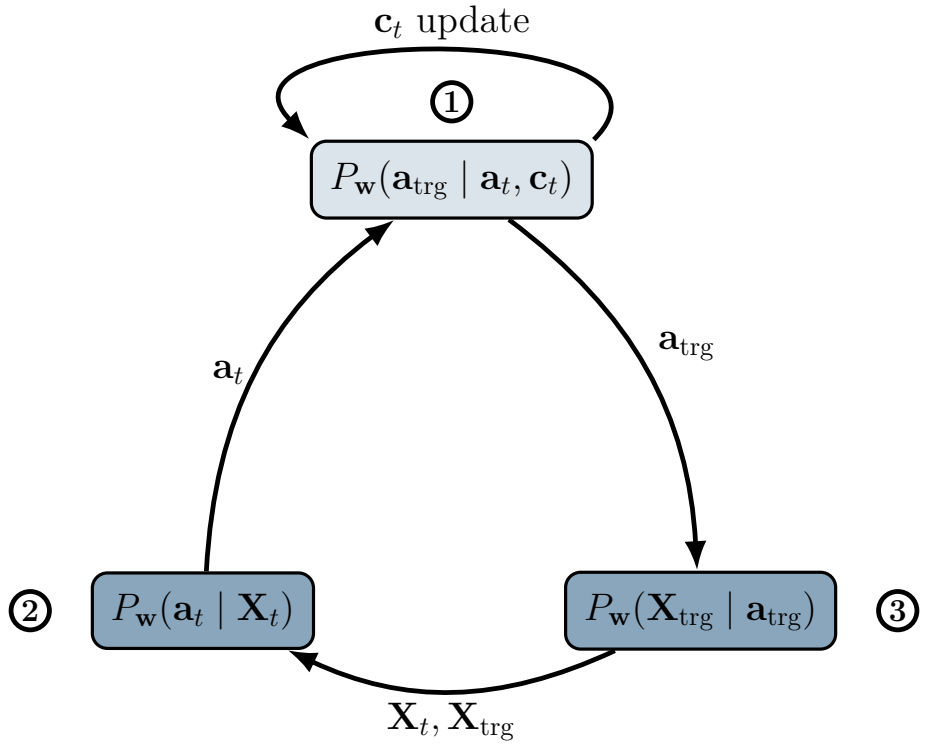


Figure 5.2: Illustration of the Unified Loop, where Affect-*understanding* ②: $P_{\mathbf{w}}(\mathbf{a}_t \mid \mathbf{X}_t)$ and Affect-*synthesis* ③: $P_{\mathbf{w}}(\mathbf{X}_{\text{trg}} \mid \mathbf{a}_{\text{trg}})$ are connected through an Affect-*control* system ①: $P_{\mathbf{w}}(\mathbf{a}_{\text{trg}} \mid \mathbf{a}_t, \mathbf{c}_t)$, forming a closed loop. A self-loop on the control module denotes iterative updates of the interaction context $\mathbf{c}_t$. Note that, between the connection linking nodes ② and ③, $\mathbf{X}_{\text{trg}}$ represents the agent's affective response, whereas $\mathbf{X}_t$ corresponds to the respective speech response of a human in the group.

While this thesis has primarily advanced the emotion-understanding module $P_{\mathbf{w}}(\mathbf{a}_t \mid \mathbf{X}_t)$ and the emotion-synthesis module $P_{\mathbf{w}}(\mathbf{X}_{\text{trg}} \mid \mathbf{a}_{\text{trg}})$, future research should turn to the emotion-control module $P_{\mathbf{w}}(\mathbf{a}_{\text{trg}} \mid \mathbf{a}_t, \mathbf{c}_t)$. This component is essential for enabling adaptive behavior in both embodied and disembodied socially intelligent agents, allowing the system to dynamically regulate its target affective state in response to the user's affect and the evolving interaction context $\mathbf{c}_t$. Prior work on adaptive virtual agents shows that adaptation at behavioral, conversational, and signal levels strongly influences user engagement and impressions [298]. Building on these insights, probabilistic formulations of the control module would enable

agents not only to select deterministic responses, but also to represent and reason over uncertainty in user expectations, affective trajectories, and context-dependent states. Such an approach could yield agents capable of flexibly adapting affective behavior in real time, aligning more closely with diverse user preferences, and ultimately supporting richer and more contextually appropriate human–agent interactions. At the same time, several challenges must be addressed: (i) balancing responsiveness with stability so that adaptations are neither erratic nor rigid, (ii) ensuring that probabilistic control models remain computationally efficient enough for real-time deployment in interactive systems, and (iii) the difficulty of reliably estimating evolving interaction contexts $\mathbf{c}_t$ in noisy, in-the-wild environments.

5.3 Reflecting on Interdisciplinary Research and Collaborations

Interdisciplinary and cross-disciplinary research involve integrating insights, methods, and perspectives from multiple fields to address complex questions that cannot be fully understood within a single discipline. While cross-disciplinary work draws on knowledge from one field to inform another, interdisciplinary research emphasizes deeper collaboration and integration across domains. A central cornerstone of this thesis is the pursuit of robust interdisciplinary research bridging computer science with the social sciences. From the perspective of a computer scientist, this has required not only engaging deeply with theories from social sciences, but also ensuring that domain knowledge from these fields directly informs the research design—shaping the formulation of research questions, the choice of modeling techniques, and the interpretation of results.

Strong interdisciplinary collaborations bring tangible benefits: they guard against construct proliferation[1], promote theory–method alignment, and ensure that concepts and terminologies are not siloed within disciplinary boundaries. Yet, while these collaborations are vital, they are also complex and fraught with challenges that must be carefully navigated.

Figure 5.3 illustrates how our three contributions are positioned along the spectrum of interdisciplinary research. Since this thesis was written from the perspective of a computer scientist within the Department of Informatics at the University of Hamburg, it is evident that all three contributions are rooted primarily in computer science but extend toward interdisciplinary integration (positioned around the middle of the spectrum). From this vantage point, the degree to which a contribution draws upon and integrates insights from social sciences and organizational psychology determines how far to the right it is positioned.

The first contribution represents a moderate degree of interdisciplinarity. The research question it addresses—label uncertainty in affect—is a challenge recognized in both social sciences and computer science (see our related contribution in [P6]). Although the problem is shared across the two domains, the solution developed here is primarily computer science–driven. This underscores a key point: interdisciplinarity does not always require solving a problem jointly across disciplines, but it can emerge when a problem of mutual interest is reframed and tackled rigorously within one field while still maintaining relevance for another.

The second contribution, in our view, achieves the most balanced and optimal degree of interdisciplinary collaboration *among the chapters of this thesis*. Crucially, it goes beyond

[1] Construct proliferation refers to the tendency in social sciences and psychology to introduce overlapping or narrowly defined constructs, which can fragment knowledge and hinder theory building.

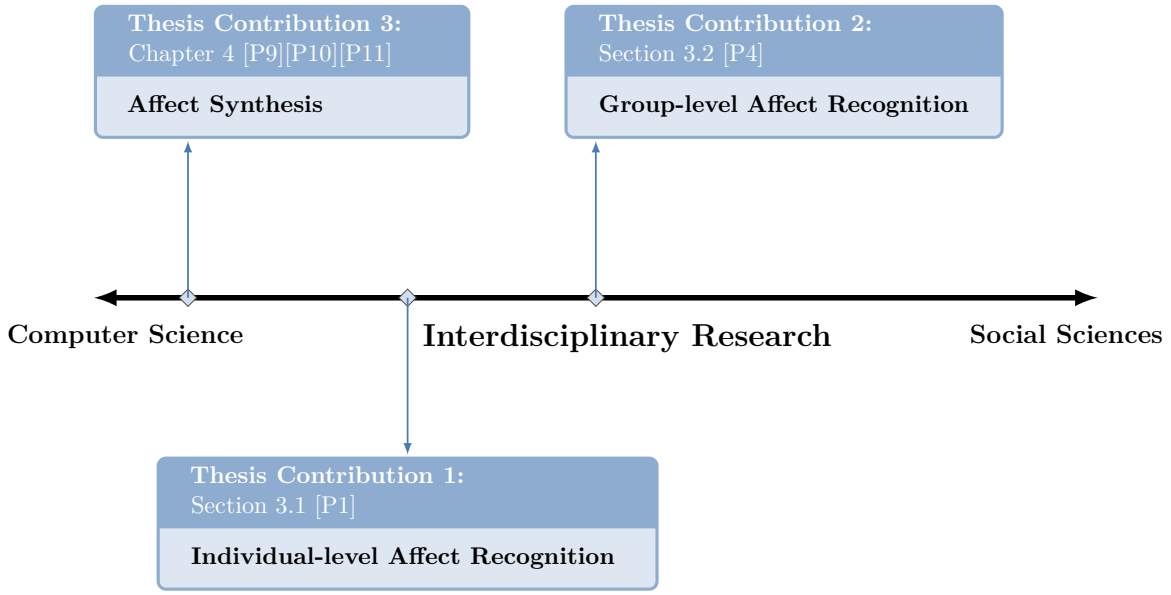


Figure 5.3: Illustration of contributions situated along the spectrum of interdisciplinary research between computer science and the social sciences.

the common producer–consumer model of interdisciplinary research, where the social sciences provide the dataset and the computer sciences primarily consume it for modeling purposes. This work not only advances computer science but also makes a direct and original contribution to organizational psychology. Importantly, it addresses the longstanding theory–method misalignment in affective computing, particularly in the context of group-level affect. By bridging this gap, it offers a process-oriented analysis of interpersonal affective convergence and divergence within group interactions—an area largely unexplored in organizational psychology. Such progress was possible only through close collaboration: organizational psychologists informed the annotation collection process, while insights from the social sciences guided the design of our graph-based modeling approach. This contribution illustrates how true interdisciplinarity emerges when methodological innovation and domain knowledge are deeply interwoven.

The third contribution, though central to advancing socially situated perception (SSP) research, is less interdisciplinary in nature, as it does not directly address a research gap within the social sciences. Nevertheless, the research design carefully avoided ignoring or contradicting established affect theories from social sciences. For instance, while most prior work relied on acted datasets and categorical affect labels, our approach was the first in this area to move toward complex, in-the-wild datasets and adopt the circumplex model of continuous affect quantification. By tackling these more challenging frontiers, this contribution maintained theoretical alignment with social sciences even while its primary impact lies within computer science.

Challenges within interdisciplinary research

Inherent orientation towards one discipline. A fundamental challenge of interdisciplinary research is that contributions are often naturally pulled toward one end of the spectrum, depending on the lead contributor's disciplinary background—whether computer science or the social sciences. Genuine interdisciplinarity requires either (i) close collaboration with

experts from the other domain, (ii) substantial effort by the lead contributor to gain sufficient expertise outside their home field, or ideally (iii) a combination of both. Each of these paths is demanding, as they involve balancing the interests, agendas, and incentives of multiple parties. Moreover, for collaborations to be sustained and meaningful, contributions must strive to address knowledge gaps that matter to *both* research communities. In this regard, the *Thesis Contribution 2* exemplifies the third pathway: the lead contributor invested efforts to acquire expertise in group affect and actively facilitated collaborations across domains. Organizational psychologists were engaged in shaping the annotation strategy and guiding the convergence–divergence analysis, while computer scientists ensured that insights from organizational psychology directly informed methodological design.

Joint efforts and conditions for co-design. The interdisciplinary co-design process highlighted above requires significant and continuous effort, which in our view depends on two factors: (i) ensuring joint benefits alongside distinct contributions for collaborators in their respective fields, and (ii) a willingness to step beyond one's disciplinary boundaries to identify potential theory–method mismatches or overlooked knowledge gaps. Mechanisms such as regular short collaboration meetings or, when possible, longer interdisciplinary workshops can provide space for reflection, alignment, and shared motivation. These forums not only help prevent theory–method misalignment but also ensure that domain-specific expertise is effectively leveraged. Ultimately, mutual benefit is best realized when interdisciplinary research aims at tangible demonstrations and applications in real-world settings—a limitation we recognize in this thesis. Another promising approach is early exposure: embedding interdisciplinary concepts (e.g., social signal processing) into undergraduate or master's-level courses can prepare students for thesis work that is naturally positioned at the intersection of disciplines.

Challenges of disciplinary perspectives. As noted, joint efforts flourish only when contributions generate value for researchers across the spectrum. Yet, this is difficult to achieve because scholars often interpret work through *disciplinary perspectives*, which may cause them to overlook challenges or knowledge gaps in the other field. This misalignment can arise at multiple levels: (i) between funding proposal writers and reviewers, (ii) between collaborators, and/or (iii) between authors and peer reviewers. A central difficulty arises from the breadth of knowledge required, as it is inherently demanding for any researcher to keep pace with advances in both fields of an interdisciplinary project. Reflecting on this thesis, we conclude that the primary responsibility for achieving interdisciplinarity rests with the lead contributor, who must actively drive collaborations, frame mutually beneficial research agendas, and integrate cross-disciplinary expertise. At the same time, reviewers of interdisciplinary work also carry a demanding role: they must remain sufficiently aware of advances in both fields to accurately assess and fairly judge contributions that are interdisciplinary in nature.

Summary Looking back across the three contributions of this thesis, my central takeaway is that interdisciplinarity is not a by-product of combining methods from different fields but a sustained, often demanding commitment to integrating ways of thinking. In my experience, the most meaningful progress occurred when I was willing to step outside my disciplinary comfort zone—engaging with theory-building in organizational psychology, questioning assumptions in affective computing, and learning to translate between two research cultures with different priorities and vocabularies. This process was not always smooth: it required negotiating differing expectations, adapting research designs in response to domain-specific constraints, and recognising when a solution that was elegant from a computer science perspective did not yet address what mattered to psychologists, and vice versa.

What I ultimately learned is that successful interdisciplinary work requires an active, deliberate effort to hold both perspectives in view simultaneously—not alternating between them, but integrating them in a way that shapes the research from the outset. When this alignment was achieved, most clearly in *Thesis Contribution 2*, the outcome was qualitatively different from what either field could have produced alone: a contribution that addressed a theoretical gap in organizational psychology while advancing methodological innovation in computer science. While this commitment to interdisciplinarity must be especially strong on the part of the lead contributor for the research to be effective, it must also be upheld by multiple stakeholders across the project lifecycle—between funding proposal writers and reviewers, among collaborators, and during peer review—to ensure that interdisciplinary work is recognised, supported, and fairly evaluated.

This, to me, is the key lesson of the interdisciplinary journey reflected in this thesis: genuine interdisciplinarity is less about balancing disciplines on a spectrum and more about designing research questions that only make sense when the strengths of both domains come together. It is this integrative perspective that I hope to carry forward beyond the PhD.

References

[1] N. Jia, X. Luo, Z. Fang, and C. Liao, "When and how artificial intelligence augments employee creativity," *Academy of Management Journal*, vol. 67, no. 1, pp. 5–32, 2024.

[2] S. Wang, Z. Sun, and Y. Chen, "Effects of higher education institutes' artificial intelligence capability on students' self-efficacy, creativity and learning performance," *Education and Information Technologies*, vol. 28, no. 5, pp. 4919–4939, 2023.

[3] T. Weber, M. Brandmaier, A. Schmidt, and S. Mayer, "Significant productivity gains through programming with large language models," *Proc. on Human-Computer Interaction*, vol. 8, 2024.

[4] X. Wei, L. Wang, L.-K. Lee, and R. Liu, "The effects of generative AI on collaborative problem-solving and team creativity performance in digital story creation: An experimental study," *International Journal of Educational Technology in Higher Education*, vol. 22, p. 23, Apr. 2025.

[5] J. Ashkinaze, J. Mendelsohn, Q. Li, C. Budak, and E. Gilbert, "How ai ideas affect the creativity, diversity, and evolution of human ideas: Evidence from a large, dynamic experiment," *CoRR*, vol. abs/2401.13481, 2024.

[6] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and vision computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[7] X. Zhou, H. Zhu, L. Mathur, *et al.*, "SOTOPIA: Interactive evaluation for social intelligence in language agents," in *Proc. Int. Conf. Learning Representations*, 2024.

[8] K. Dautenhahn, "Getting to know each other—artificial social intelligence for autonomous robots," *Robotics and autonomous systems*, vol. 16, no. 2-4, pp. 333–356, 1995.

[9] T. J. Wiltshire, E. J. Lobato, J. Velez, F. Jentsch, and S. M. Fiore, "An interdisciplinary taxonomy of social cues and signals in the service of engineering robotic social intelligence," in *Unmanned Systems Technology XVI*, SPIE, vol. 9084, 2014, pp. 124–138.

[10] K. Dautenhahn, *A paradigm shift in artificial intelligence: why social intelligence matters in the design and development of robots with human-like intelligence*. Springer, 2007.

[11] A. Pentland, "Social signal processing [exploratory dsp]," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 108–111, 2007.

[12] C. Raman, "Towards artificial social intelligence in the wild: Sensing, synthesizing, modeling, and perceiving nonverbal social human behavior," Dissertation, Delft University of Technology, 2023.

[13] J. B. Schmutz, N. Outland, S. Kerstan, E. Georganta, and A.-S. Ulfert, "Ai-teaming: Redefining collaboration in the digital era," *Current Opinion in Psychology*, vol. 58, p. 101 837, 2024.

[14] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science robotics*, vol. 3, no. 21, 2018.

[15] S. Rasouli, G. Gupta, E. Nilsen, and K. Dautenhahn, "Potential applications of social robots in robot-assisted interventions for social anxiety," *International Journal of Social Robotics*, vol. 14, no. 5, pp. 1–32, 2022.

[16] N. Hirvonen, V. Jylhä, Y. Lao, and S. Larsson, "Artificial intelligence in the information ecosystem: Affordances for everyday information seeking," *Journal of the Association for Information Science and Technology*, vol. 75, no. 10, pp. 1152–1165, 2024.

[17] S. G. Barsade and A. P. Knight, "Group affect," *Annu. Rev. Organ. Psychol. Organ. Behav.*, vol. 2, no. 1, pp. 21–46, 2015.

[18] S. Hinsen, P. Hofmann, J. Jöhnk, and N. Urbach, "How can organizations design purposeful human-ai interactions: A practical perspective from existing use cases and interviews," in *Proc. of the 55th Hawaii Int. Conf. on System Sciences*, Jan. 2022.

[19] E. Georganta, S. Stracke, T. A. O'Neill, *et al.*, "Looking forward: Introducing artificial intelligence in teams," *Academy of Management Proceedings*, vol. 2022, no. 1, p. 14 144, 2022.

[20] N. Lehmann-Willenbrock, S. G. Rogelberg, J. A. Allen, and J. E. Kello, "The critical importance of meetings to leader and organizational success: Evidence-based insights and implications for key stakeholders," *Organizational Dynamics*, vol. 47, no. 1, p. 32, 2017.

[21] M. A. West, *Effective teamwork: Practical lessons from organizational research*. John Wiley & Sons, 2012.

[22] S. Kauffeld and N. Lehmann-Willenbrock, "Meetings matter: Effects of team meetings on team and organizational success," *Small group research*, vol. 43, no. 2, pp. 130–158, 2012.

[23] S. W. Kozlowski, "Enhancing the effectiveness of work groups and teams: A reflection," *Perspectives on Psychological Science*, vol. 13, no. 2, pp. 205–212, 2018.

[24] N. Churamani, P. Barros, H. Gunes, and S. Wermter, "Affect-driven learning of robot behaviour for collaborative human-robot interactions," *Frontiers in Robotics and AI*, vol. Volume 9 - 2022, 2022.

[25] M. Houtti, M. Zhou, L. Terveen, and S. Chancellor, "Observe, ask, intervene: Designing ai agents for more inclusive meetings," in *Proc. of the Conf. on Human Factors in Computing Systems*, ser. CHI '25, New York, USA: ACM, 2025.

[26] K. R. Scherer, "What are emotions? and how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[27] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant.," *Journal of personality and social psychology*, vol. 76, no. 5, p. 805, 1999.

[28] J. A. Russell, "Core affect and the psychological construction of emotion.," *Psychological review*, vol. 110, no. 1, p. 145, 2003.

[29] S. G. Barsade and D. E. Gibson, "Why does affect matter in organizations?" *Academy of management perspectives*, vol. 21, no. 1, pp. 36–59, 2007.

[30] S. G. Barsade and D. E. Gibson, "Group affect: Its influence on individual and group outcomes," *Current Directions in Psychological Science*, vol. 21, no. 2, pp. 119–123, 2012.

[31] Z. Lei and N. Lehmann-Willenbrock, "Dynamic affect in team meetings: An interpersonal construct embedded in dynamic interaction processes," in *The Cambridge handbook of meeting science*, Cambridge University Press, 2015, pp. 456–480.

[32] A. P. Brief and H. M. Weiss, "Organizational behavior: Affect in the workplace," *Annual Review of Psychology*, vol. 53, no. 1, 2002.

[33] R. W. Picard, *Affective computing*. Cambridge, MA, USA: MIT Press, 1997.

[34] B. W. Schuller, "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.

[35] S. Alisamir and F. Ringeval, "On the evolution of speech representations for affective computing: A brief history and critical overview," *IEEE Signal Proc., Magazine*, vol. 38, pp. 12–21, 2021.

[36] G. Trigeorgis, F. Ringeval, R. Brueckner, *et al.*, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, 2016.

[37] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, *et al.*, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Tran. on Pattern Analysis and Machine Int.*, 2023.

[38] D. Diatlova, A. Udalov, V. Shutov, and E. Spirin, "Adapting wavlm for speech emotion recognition," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024, pp. 303–308.

[39] K. R. Scherer and H. Ellgring, "Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal?" *Emotion*, vol. 7, no. 1, pp. 113–130, 2007.

[40] C. Raman, N. Raj Prabhu, and H. Hung, "Perceived conversation quality in spontaneous interactions," *IEEE Tran. on Affective Comp.*, vol. 14, no. 4, pp. 2901–2912, 2023.

[41] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[42] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Tran. on Affective Comp.*, vol. 10, no. 4, pp. 471–483, Dec. 2019.

[43] E. A. Veltmeijer, C. Gerritsen, and K. V. Hindriks, "Automatic emotion recognition for groups: A review," *IEEE Tran. on Affective Computing (TAFFC)*, vol. 14, no. 1, pp. 89–107, 2021.

[44] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual and spontaneous expressions," in *Proc. of the Int. Conf. on Multimodal Interfaces*, 2007, pp. 126–133.

[45] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The ambiguous world of emotion representation," *arXiv preprint arXiv:1909.00360*, 2019.

[46] B. Dudzik, T. M. Hrkalovic, C. Hao, C. Raman, and M. Tsfasman, "Indeterminacy in affective computing: Considering meaning and context in data collection practices," in *IEEE Int. Conf. on Affective Comp. and Intelligent Interaction Workshops and Demos (ACIIW)*, 2024, pp. 181–185.

[47] K. P. Murphy, *Machine learning : a probabilistic perspective*. Cambridge, USA: MIT Press, 2012.

[48] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. Cambridge, MA: MIT Press, 2022.

[49] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe, "The more the merrier: Analysing the affect of a group of people in images," in *Int. Conf. on Automatic Face and Gesture Recognition*, IEEE, 2015.

[50] X. Huang, J. Xu, W. Zheng, Q. Mao, and A. Dhall, "A survey of deep learning for group-level emotion recognition," *CoRR*, 2024.

[51] S. G. Barsade, "The ripple effect: Emotional contagion and its influence on group behavior," *Administrative science quarterly*, vol. 47, no. 4, pp. 644–675, 2002.

[52] S. Arora, K.-W. Chang, C.-M. Chien, *et al.*, "On the landscape of spoken language models: A comprehensive survey," *arXiv preprint arXiv:2504.08528*, 2025.

[53] T. A. Nguyen, E. Kharitonov, J. Copet, *et al.*, "Generative spoken dialogue language modeling," *Tran. of the Association for Computational Linguistics*, vol. 11, pp. 250–266, Mar. 2023.

[54] S. Amiriparian, B. W. Schuller, N. Asghar, H. Zen, and F. Burkhardt, "Guest editorial: Special issue on affective speech and language synthesis, generation, and conversion," *IEEE Tran. on Affective Computing (TAFFC)*, 2023.

[55] A. Triantafyllopoulos, B. W. Schuller, G. İymen, *et al.*, "An overview of affective speech synthesis and conversion in the deep learning era," *Proc. of the IEEE*, 2023.

[56] K. Zhou, "Emotion modelling for speech generation," Available at `https://scholarbank.nus.edu.sg/handle/10635/243782`, PhD thesis, National University of Singapore, 2022.

[57] A. N. Salman, Z. Du, S. S. Chandra, İ. R. Ülgen, C. Busso, and B. Sisman, "Towards naturalistic voice conversion: Naturalvoices dataset with an automatic processing pipeline," in *Proc. Interspeech*, 2024, pp. 4358–4362.

[58] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[59] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," *Proc. Neural Inf. Process. Syst.*, vol. 28, 2015.

[60] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[61] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *Advances in Neural Inf. Proc. Sys., NeurIPS*, Vancouver, Canada, 2020.

[62] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, " High-Resolution Image Synthesis with Latent Diffusion Models," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Los Alamitos, CA, USA, Jun. 2022, pp. 10 674–10 685.

[63] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Proc. Neural Inf. Process. Syst.*, vol. 33, 2020.

[64] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *Proc. Int. Conf. Machine Learning*, PMLR, 2021.

[65] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. Hashimoto, "Diffusion-LM improves controllable text generation," in *Proc. Neural Inf. Process. Syst.*, 2022.

[66] J. R. Hackman and N. Katz, "Group behavior and performance," *Handbook of social psychology*, vol. 2, pp. 1208–1251, 2010.

[67] E. Goffman, *Behavior in public places*. Simon and Schuster, 2008.

[68] A. Kendon, *Conducting interaction: Patterns of behavior in focused encounters*. Cambridge University Press, 1990, vol. 7.

[69] J. E. McGrath and I. Altman, "Small group research: A synthesis and critique of the field.," 1966.

[70] J. R. Hackman and C. G. Morris, "Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration," *Advances in experimental social psychology*, vol. 8, pp. 45–99, 1975.

[71] S. G. Barsade and D. E. Gibson, "Group emotion: A view from top and bottom," in *Composition* (Research on managing groups and teams), Research on managing groups and teams. Elsevier Science/JAI Press, 1998, vol. 1, pp. 81–102.

[72] J. A. Allen, N. Lehmann-Willenbrock, and S. J. Sands, "Meetings as a positive boost? How and when meeting satisfaction impacts employee empowerment," *Journal of Business Research*, vol. 69, no. 10, pp. 4340–4347, 2016.

[73] N. Lehmann-Willenbrock, "Dynamic interpersonal processes at work: Taking social interactions seriously," *Annual Review of Organizational Psychology and Organizational Behavior*, 2024.

[74] P. Jarzabkowski and D. Seidl, "The role of meetings in the social practice of strategy," *Organization studies*, vol. 29, no. 11, pp. 1391–1426, 2008.

[75] Atlassian, *Time wasting at work: Meetings, emails, and distractions*, `https://www.atlassian.com/time-wasting-at-work-infographic`, 2022.

[76] Fellow.app, *Meeting statistics: How much time do we spend in meetings?* `https://fellow.app/blog/meetings-statistics-how-many-hours-do-we-spend-in-meetings`, 2024.

[77] Flowtrace, *Average time in meetings and its impact*, `https://www.flowtrace.co/collaboration-blog/average-time-in-meetings-its-impact`, 2024.

[78] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone, "Evidence for a collective intelligence factor in the performance of human groups," *Science*, vol. 330, no. 6004, pp. 686–688, 2010.

[79] A. Pentland, "The new science of building great teams," *Harvard Business Review*, vol. 90, no. 4, pp. 60–69, 2012.

[80] J. A. Allen, N. Lehmann-Willenbrock, and S. G. Rogelberg, "Let's get this meeting started: Meeting lateness and actual meeting outcomes," *Journal of Organizational Behavior*, vol. 36, no. 4, pp. 687–709, 2015.

[81] H. H. Clark and S. E. Brennan, "Grounding in communication," in *Perspectives on socially shared cognition*, L. B. Resnick, J. M. Levine, and S. D. Teasley, Eds., American Psychological Association, 1991, pp. 127–149.

[82] E. W. Morrison and F. J. Milliken, "Speaking up, remaining silent: The dynamics of voice and silence in organizations," *Academy of Management Review*, vol. 25, no. 4, pp. 706–725, 2000.

[83] J. O'Neill, A. Woodruff, and J. Forlizzi, "Smart meeting rooms: Toward automated understanding of nonverbal participants' engagement," in *Proceedings of the ACM on Computer Supported Cooperative Work (CSCW)*, ACM, 2018.

[84] R. A. Guzzo and G. P. Shea, "Productivity and team effectiveness: A review of group-level interventions," *Research in organizational behavior*, vol. 7, pp. 269–331, 1985.

[85] D. R. Ilgen, J. R. Hollenbeck, M. Johnson, and D. Jundt, "Teams in organizations: From input-process-output models to imoi models," *Annual Review of Psychology*, vol. 56, pp. 517–543, 2005.

[86] T. A. Judge and R. F. Piccolo, "Transformational and transactional leadership: A meta-analytic test of their relative validity.," *Journal of applied psychology*, vol. 89, no. 5, p. 755, 2004.

[87] M. A. Campion, G. J. Medsker, and A. C. Higgs, "Relations between work group characteristics and effectiveness: Implications for designing effective work groups," *Personnel Psychology*, vol. 46, no. 4, pp. 823–850, 1993.

[88] J. R. Hackman, *Leading Teams: Setting the Stage for Great Performances*. Harvard Business Press, 2002.

[89] A. Knight and N. Eisenkraft, "Positive is usually good, negative is not always bad: The effects of group affect on social integration and task performance," *Journal of Applied Psychology*, vol. 100, pp. 1214–1227, Dec. 2014.

[90] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis.," *Psychological bulletin*, vol. 111, no. 2, p. 256, 1992.

[91] Z. Lei and N. Lehmann-Willenbrock, "Affect in meetings: An interpersonal construct in dynamic interaction processes," in *The Cambridge handbook of meeting science*, 2015, pp. 456–482.

[92] S. W. Kozlowski, "Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations," *Organizational Psychology Review*, vol. 5, no. 4, pp. 270–299, 2015.

[93] N. Lehmann-Willenbrock and H. Hung, "A multimodal social signal processing approach to team interactions," *Organizational Research Methods*, 2024.

[94] V. Begemann, N. Lehmann-Willenbrock, and M. Stein, "Peeling away the layers of workplace gossip: A framework, review, and future research agenda to study workplace gossip as a dynamic and complex behavior," *Merits*, vol. 3, no. 2, pp. 297–317, 2023.

[95] N. Lehmann-Willenbrock, "Dynamic interpersonal processes at work: Taking social interactions seriously," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 12, 2024.

[96] Y. Zhao, T. Liu, X. Han, and H. Gui, "Team decision-making interaction and performance: A behavioral process-based relationship study," *Small Group Research*, vol. 55, no. 6, pp. 919–952, 2024.

[97] M. Grabowski, N. Lehmann-Willenbrock, S. Rings, A. Blanchard, and F. Steinicke, "Group dynamics in the metaverse: A conceptual framework and first empirical insights," *Small Group Research*, vol. 55, no. 5, pp. 763–804, 2024.

[98] S. M. Fingas, C. Busch, R. Dreyer, and N. Lehmann-Willenbrock, "Zooming in: Identifying fine-grained verbal dynamics that influence coachees' self-regulation statements during copreneur coaching sessions," *Journal of Occupational and Organizational Psychology*, vol. 98, no. 2, 2025.

[99] K. Oatley, D. Keltner, and J. M. Jenkins, *Understanding emotions*. Blackwell publishing, 2006.

[100] G. A. Van Kleef, "How emotions regulate social life: The emotions as social information (easi) model," *Current directions in psychological science*, vol. 18, no. 3, pp. 184–188, 2009.

[101] G. A. van Kleef, M. W. Heerdink, and A. C. Homan, "Emotional influence in groups: The dynamic nexus of affect, cognition, and behavior," *Current Opinion in Psychology*, vol. 17, pp. 156–161, 2017.

[102] K. R. Scherer, "Appraisal considered as a process of multilevel sequential checking," *Appraisal processes in emotion: Theory, methods, research*, vol. 92, no. 120, p. 57, 2001.

[103] R. S. Lazarus, *Emotion and adaptation* (Emotion and adaptation). New York, NY, US: Oxford University Press, 1991, pp. xiii, 557.

[104] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[105] C. J. Beedie, P. C. Terry, and A. M. Lane, "Distinctions between emotion and mood," *Cognition and Emotion*, vol. 19, no. 6, pp. 847–878, 2005.

[106] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[107] K. R. Scherer, "Toward a dynamic theory of emotion: The component process model of affective states," *Geneva studies in Emotion and Communication*, vol. 1, pp. 1–98, 1987.

[108] N. Raj Prabhu, C. Raman, and H. Hung, "Defining and Quantifying Conversation Quality in Spontaneous Interactions," in *Comp. Pub. of ACM Int. Conf. on Multimodal Interaction*, Sep. 2020.

[109] A. S. Gabriel, M. A. Daniels, J. M. Diefendorff, and G. J. Greguras, "Emotional labor actors: A latent profile analysis of emotional labor strategies," *Journal of Applied Psychology*, vol. 100, no. 3, 2015.

[110] D. Geddes and D. Lindebaum, "Unpacking the 'why' behind strategic emotion expression at work: A narrative review and proposed taxonomy," *European Management Journal*, 2020.

[111] F. Liu and S. Maitlis, "Emotional dynamics and strategizing processes: A study of strategic conversations in top team meetings," *Journal of management studies*, vol. 51, no. 2, pp. 202–234, 2014.

[112] T. M. Glomb and M. J. Tews, "Emotional labor: A conceptualization and scale development," *Journal of Vocational Behavior*, vol. 64, no. 1, pp. 1–23, 2004.

[113] J. M. Hektner, J. A. Schmidt, and M. Csikszentmihalyi, *Experience sampling method: Measuring the quality of everyday life*. Sage, 2007.

[114] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[115] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Interspeech*, Shanghai, China, Oct. 2020.

[116] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[117] J. A. Allen, C. Fisher, M. Chetouani, *et al.*, "Comparing social science and computer science workflow processes for studying group interactions," *Small Group Research*, vol. 48, no. 5, pp. 568–590, 2017.

[118] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal communication in human interaction*. Thomson Wadsworth, 1972.

[119] V. P. Richmond, "Nonverbal behavior in interpersonal relations," *(No Title)*, p. 366, 2008.

[120] M. S. Bedmutha, A. Tsedenbal, K. Tobar, *et al.*, "Conversense: An automated approach to assess patient-provider interactions using social signals," in *Proc. of the Conf. on Human Factors in Computing Systems (CHI)*, Honolulu, HI, USA: Association for Computing Machinery (ACM), 2024.

[121] R. W. Picard, *Affective computing*. Cambridge, MA, USA: MIT Press, 1997.

[122] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Tran. on Affective Computing (TAFFC)*, vol. 1, no. 1, pp. 18–37, 2010.

[123] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," in *IEEE/CVF Winter Conf. on App. of Computer Vision*, IEEE, 2016.

[124] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, 2018.

[125] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," 5, vol. 61, ACM New York, NY, USA, 2018, pp. 90–99.

[126] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis: A review," *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 35–55, 2019.

[127] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: A review," in *Int. colloquium on signal processing and its applications*, IEEE, 2011, pp. 410–415.

[128] A. Halbig and M. E. Latoschik, "A systematic review of physiological measurements, factors, methods, and applications in virtual reality," *Frontiers in Virtual Reality*, vol. 2, p. 694 567, 2021.

[129] C. Darwin, *The Expression of the Emotions in Man and Animals*. John Murray, 1872.

[130] P. Lieberman, J. T. Laitman, J. S. Reidenberg, and P. J. Gannon, "The anatomy, physiology, acoustics and perception of speech: Essential elements in analysis of the evolution of human speech," *Journal of Human Evolution*, vol. 23, no. 6, pp. 447–467, 1992.

[131] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[132] M. J. Traxler, *Introduction to Psycholinguistics: Understanding Language Science*. Wiley-Blackwell, 2012.

[133] C. Tang, W. Yu, G. Sun, *et al.*, "Salmonn: Towards generic hearing abilities for large language models," in *Proc. Int. Conf. Learning Representations*, 2024.

[134] Q. Fang, S. Guo, Y. Zhou, Z. Ma, S. Zhang, and Y. Feng, "LLaMA-omni: Seamless speech interaction with large language models," in *Proc. Int. Conf. Learning Representations*, 2025.

[135] K. Chen, Y. Gou, R. Huang, *et al.*, "Emova: Empowering language models to see, hear and speak with vivid emotions," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Jun. 2025, pp. 5455–5466.

[136] Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro, "Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities," in *Proc. Int. Conf. Machine Learning*, vol. 235, PMLR, Jul. 2024, pp. 25 125–25 148.

[137] I. McCowan, J. Carletta, W. Kraaij, *et al.*, "The ami meeting corpus," in *Proc. of the 5th Int. Conf. on Methods and Techniques in Behavioral Research*, vol. 88, 2005, p. 100.

[138] L. Maman, E. Ceccaldi, N. Lehmann-Willenbrock, *et al.*, "GAME-ON: A Multimodal Dataset for Cohesion and Group Analysis," en, *IEEE Access*, vol. 8, pp. 124 185–124 203, 2020.

[139] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, Jan. 2017.

[140] L. Stappen, A. Baird, L. Schumann, and B. Schuller, "The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements," *IEEE Tran. on Affective Computing (TAFFC)*, vol. 14, no. 2, pp. 1334–1350, 2021.

[141] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Tran. on Affective Computing (TAFFC)*, vol. 3, no. 1, pp. 5–17, 2011.

[142] D. Kollias and S. Zafeiriou, " Analysing Affective Behavior in the second ABAW2 Competition," pp. 3645–3653, Oct. 2021.

[143] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, e0196391, 2018.

[144] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," in *IEEE Tran. on Affective Computing (TAFFC)*, 2014.

[145] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of ACL*, 2018.

[146] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Tran. on Affective Computing (TAFFC)*, vol. 12, no. 2, pp. 479–493, 2021.

[147] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," in *IEEE Tran. on Affective Computing (TAFFC)*, vol. 10, 2019, pp. 18–31.

[148] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of ACL*, 2019.

[149] J. Comas, A. J. Vera, X. Vives, E. De Filippi, A. Pereda, and F. Sukno, "Cast-phys: Contactless affective states through physiological signals database," *arXiv preprint arXiv:2507.06080*, 2025.

[150] M. Lee, A. Shomanov, B. Begim, *et al.*, "Eav: Eeg-audio-video dataset for emotion recognition in conversational contexts," *Scientific Data*, vol. 11, 2024.

[151] P. Yang, N. Liu, X. Liu, *et al.*, "A multimodal dataset for mixed emotion recognition," *Scientific Data*, vol. 11, Aug. 2024.

[152] X. Huang, S. Zhu, Z. Wang, Y. He, H. Jin, and Z. Liu, "Eva-med: An enhanced valence-arousal multimodal emotion dataset for emotion recognition," 2025.

[153] Y. Li, W. Gan, K. Lu, D. Jiang, and R. Jain, "AVES: An audio-visual emotion stream dataset for temporal emotion detection," *IEEE Tran. on Affective Computing (TAFFC)*, vol. 16, no. 1, pp. 438–450, 2025.

[154] H. Sun, X. Wang, J. Zhao, *et al.*, "EmotionTalk: An interactive chinese multimodal emotion dataset with rich annotations," *arXiv preprint arXiv:2505.23018*, 2025.

[155] J. Kossaifi, R. Walecki, Y. Panagakis, *et al.*, "SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2019.

[156] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.

[157] M. Tsfasman, K. Fenech, M. Tarvirdians, A. Lorincz, C. Jonker, and C. Oertel, "Towards creating a conversational memory for long-term meeting support: Predicting memorable moments in multi-party conversations through eye-gaze," in *ACM Int. Conf. on Multimodal Interaction*, Bengaluru, India, 2022, pp. 94–104.

[158] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung, "The MatchNMingle dataset: A novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates," *IEEE Tran. on Affective Comp.*, vol. 12, no. 1, pp. 113–130, 2021.

[159] A. Dhall, J. Joshi, I. Radwan, and R. Goecke, "Finding happiest moments in a social context," in *Asian Conference on Computer Vision*, Springer, 2012, pp. 613–626.

[160] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "Emotiw 2016: Video and group-level emotion recognition challenges," in *ACM Int. Conf. on Multimodal Interaction*, Tokyo, Japan: Association for Computing Machinery, 2016, pp. 427–432.

[161] W. Mou, O. Celiktutan, and H. Gunes, "Group-level arousal and valence recognition in static images: Face, body and context," in *11th International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, IEEE, vol. 5, 2015, pp. 1–6.

[162] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "Emotiw 2018: Audio-video, student engagement and group-level affect prediction," in *ACM Int. Conf. on Multimodal Interaction*, 2018.

[163] S. Ghosh, A. Dhall, N. Sebe, and T. Gedeon, "Predicting group cohesiveness in images," in *2019 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2019, pp. 1–8.

[164] W. Mou, H. Gunes, and I. Patras, "Alone versus in-a-group: A multi-modal framework for automatic affect recognition," *ACM Tran. on Multimedia Comp., Comm., and Appl.*, vol. 15, no. 2, 2019.

[165] A. Dhall, G. Sharma, R. Goecke, and T. Gedeon, "Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 784–789.

[166] X. Guo, L. Polania, B. Zhu, C. Boncelet, and K. Barner, "Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases," in *IEEE/CVF Winter Conf. on App. of Computer Vision*, 2020, pp. 2921–2930.

[167] Y. Wang, S. Zhou, Y. Liu, K. Wang, F. Fang, and H. Qian, "Congnn: Context-consistent cross-graph neural network for group emotion recognition in the wild," *Information Sciences*, 2022.

[168] K. G. Quach, N. Le, C. N. Duong, I. Jalata, K. Roy, and K. Luu, "Non-volume preserving-based fusion to group-level emotion recognition on crowd videos," *Pattern Recognition*, vol. 128, p. 108 646, 2022.

[169] S. Madan, S. Ghosh, L. R. Sookha, *et al.*, "MIP-GAF: A mllm-annotated benchmark for important person localization and group context understanding," in *IEEE/CVF Winter Conf. on App. of Computer Vision*, 2025.

[170] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[171] F. Eyben, K. R. Scherer, B. W. Schuller, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Tran. on Affective Comp.*, vol. 7, no. 02, pp. 190–202, Apr. 2016.

[172] B. Schuller, S. Steidl, A. Batliner, *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech*, 2013.

[173] P. Tzirakis, "End2you: Multimodal profiling by end-to-end learning and applications," in *Proc. of the 1st Int. on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop*, 2020.

[174] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Calgary, Apr. 2018.

[175] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Information Fusion*, vol. 68, pp. 46–53, 2021.

[176] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[177] S. Hershey, S. Chaudhuri, D. P. Ellis, *et al.*, "Cnn architectures for large-scale audio classification," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, 2017, pp. 131–135.

[178] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Inf. Proc. Sys., NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[179] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE Tran. on on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[180] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.

[181] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," *IEEE Transactions on Affective Computing*, pp. 1–17, Jun. 2022.

[182] A. Triantafyllopoulos, A. Batliner, and B. W. Schuller, *Charting 15 years of progress in deep learning for speech emotion recognition: A replication study*, 2025.

[183] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2697–2709, 2020.

[184] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[185] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[186] K. Sridhar and C. Busso, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Barcelona, Spain, May 2020.

[187] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, 2008.

[188] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Jan. 2005, pp. 381–385.

[189] J. Han, Z. Zhang, Z. Ren, and B. Schuller, "Exploring perception uncertainty for emotion recognition in dyadic conversation and music listening," *Cognitive Computation*, vol. 13, Mar. 2021.

[190] H.-C. Chou and C.-C. Lee, "Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019, pp. 5886–5890.

[191] K. Sridhar, W.-C. Lin, and C. Busso, "Generative approach using soft-labels to learn uncertainty in predicting emotional attributes," in *IEEE Int. Conf. on Affective Comp. and Intelligent Interaction*, Virtual Event, Oct. 2021, pp. 1–8.

[192] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, pp. 120–136, 2013.

[193] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proc., of the 25th ACM Int. Conf. on Multimedia*, Mountain View, USA, Oct. 2017.

[194] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, “"of all things the measure is man" automatic classification of emotions and inter-labeler consistency,” in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Philadelphia, USA, 2005.

[195] N. M. Foteinopoulou, C. Tzelepis, and I. Patras, “Estimating continuous affect with label uncertainty,” in *IEEE Int. Conf. on Affective Comp. and Intelligent Interaction*, Virtual Event, Oct. 2021.

[196] G. Rizos and B. Schuller, “Modelling sample informativeness for deep affective computing,” in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Brighton, UK, May 2019.

[197] H.-C. Chou, W.-C. Lin, C.-C. Lee, and C. Busso, “Exploiting annotators’ typed description of emotion perception to maximize utilization of ratings for speech emotion recognition,” in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Singapore, Jan. 2022.

[198] R. Lotfian and C. Busso, “Formulating emotion perception as a probabilistic model with application to categorical emotion classification,” in *IEEE Int. Conf. on Affective Comp. and Intelligent Interaction (ACII)*, IEEE, 2017, pp. 415–420.

[199] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?” In *Advances in Neural Inf. Proc. Sys., NeurIPS*, vol. 30, Dec. 2017.

[200] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, “Weight uncertainty in neural network,” in *Int. Conf. Machine Learning (ICML)*, Lille, France, Jul. 2015.

[201] R. Zheng, S. Zhang, L. Liu, Y. Luo, and M. Sun, “Uncertainty in Bayesian deep label distribution learning,” *Applied Soft Computing*, vol. 101, Mar. 2021.

[202] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in *Int. Conf. Machine Learning (ICML), New York City, NY, USA*, Jun. 2016.

[203] M. K. Tellamekala, S. Amiriparian, B. W. Schuller, E. André, T. Giesbrecht, and M. Valstar, “Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 2, pp. 805–822, 2024.

[204] H. Fang, T. Peer, S. Wermter, and T. Gerkmann, “Integrating statistical uncertainty into neural network-based speech enhancement,” in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, Singapore, Jan. 2022.

[205] H. Fang, D. Becker, S. Wermter, and T. Gerkmann, “Integrating uncertainty into neural network-based speech enhancement,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1587–1600, 2023.

[206] E. Hüllermeier and W. Waegeman, “Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods,” *Machine learning*, vol. 110, no. 3, pp. 457–506, 2021.

[207] X. Geng, “Label distribution learning,” *Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.

[208] N. Raj Prabhu, G. Carbajal, N. Lehmann-Willenbrock, and T. Gerkmann, "End-to-end label uncertainty modeling for speech-based arousal recognition using Bayesian neural networks," in *Interspeech*, Incheon, Korea, Sep. 2022.

[209] S. Kotz and S. Nadarajah, *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.

[210] J. Kossaifi, R. Walecki, Y. Panagakis, *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Trans., on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 1022–1040, 2019.

[211] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, Nov. 2009.

[212] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability & statistics for engineers and scientists*. Pearson Education, 2007.

[213] C. Villa and F. J. Rubio, "Objective priors for the number of degrees of freedom of a multivariate t distribution and the t-copula," *Computational Statistics & Data Analysis*, vol. 124, pp. 197–219, 2018.

[214] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.

[215] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," in (5), 3rd ed., 5. MIT Press, Jul. 2016, vol. 4, ch. 3, pp. 51–77.

[216] C. Villa and S. G. Walker, "Objective prior for the number of degrees of freedom of at distribution," *Bayesian Analysis*, vol. 9, no. 1, pp. 197–220, 2014.

[217] S. W. Kozlowski and D. R. Ilgen, "Enhancing the effectiveness of work groups and teams," *Psychological science in the public interest*, vol. 7, no. 3, pp. 77–124, 2006.

[218] A. L. Collins, S. A. Lawrence, A. C. Troth, and P. J. Jordan, "Group affective tone: A review and future research directions," *Journal of Organizational Behavior*, vol. 34, no. S1, S43–S62, 2013.

[219] C. Jones, S. Volet, and D. Pino-Pasternak, "Observational research in face-to-face small groupwork: Capturing affect as socio-dynamic interpersonal phenomena," *Small Group Research*, vol. 52, no. 3, pp. 341–376, 2021.

[220] G. Sharma, A. Dhall, and J. Cai, "Audio-visual automatic group affect analysis," *IEEE Tran. on Affective Comp.*, vol. 14, no. 2, 2021.

[221] V. B. Hinsz and L. Bui, "Socially shared affect: Shared affect, affect sharing, and affective processing in groups.," *Group Dynamics: Theory, Research, and Practice*, vol. 27, no. 4, p. 229, 2023.

[222] F. Walter and H. Bruch, "The positive group affect spiral: A dynamic model of the emergence of positive affective similarity in work groups," *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, vol. 29, no. 2, pp. 239–261, 2008.

[223] S. Hareli and A. Rafaeli, "Emotion cycles: On the social influence of emotion in organizations," *Research in organizational behavior*, vol. 28, pp. 35–59, 2008.

[224] S. G. Barsade, "The ripple effect: Emotional contagion and its influence on group behavior," *Administrative Science Quarterly*, vol. 47, no. 4, pp. 644–675, 2002.

[225] Z. Lei and N. Lehmann-Willenbrock, "Contagious peers in teams: Peer affective influence on individual emotions and performance," *Academy of Management Proceedings*, vol. 2014, Oct. 2014.

[226] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *ACM Int. Conf. on Multimodal Interaction*, 2017.

[227] X. Huang, A. Dhall, R. Goecke, M. Pietikäinen, and G. Zhao, "Multimodal framework for analyzing the affect of a group of people," *Trans. on Multimedia*, vol. 20, no. 10, pp. 2706–2721, 2018.

[228] X. Wang, D. Zhang, and D.-J. Lee, "Implementing the affective mechanism for group emotion recognition with a new graph convolutional network architecture," *IEEE Tran. on Affective Comp.*, 2023.

[229] M. Tsfasman, B. Dudzik, K. Fenech, A. Lorincz, C. M. Jonker, and C. Oertel, "Introducing MeMo: A multimodal dataset for memory modelling in multiparty conversations," *Under Review*, 2024, `https://arxiv.org/abs/2409.13715`.

[230] S. Barsade and D. Gibson, "Group affect," *Current Directions in Psychological Science*, vol. 21, pp. 119–123, Mar. 2012.

[231] X. Wang, T. Chen, and D. Zhang, "A spatial-temporal graph convolutional network for video-based group emotion recognition," in *Int. Conf. on Pattern Recognition*, Springer, 2024, pp. 339–354.

[232] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Proc. Neural Inf. Process. Syst.*, vol. 30, 2017.

[233] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Int. Conf. on Learning Representations*, 2017.

[234] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *IEEE Tran. on Affective Comp.*, vol. 3, no. 3, pp. 349–365, 2012.

[235] M. C. Nanninga, Y. Zhang, N. Lehmann-Willenbrock, Z. Szlávik, and H. Hung, "Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry," in *ACM Int. Conf. on Multimodal Interaction*, 2017, pp. 206–215.

[236] M. Stel and R. Vonk, "Mimicry in social interaction: Benefits for mimickers, mimickees, and their interaction," *British journal of psychology (London, England : 1953)*, vol. 101, pp. 311–23, Aug. 2009.

[237] J. L. Hagad, R. Legaspi, M. Numao, and M. Suarez, "Predicting levels of rapport in dyadic interactions through automatic detection of posture and posture congruence," in *International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing*, IEEE, 2011, pp. 613–616.

[238] F. Bernieri and R. Rosenthal, "Interpersonal coordination: Behavior matching and interactional synchrony, in; feldman, rs, and rimé, b," *Studies in emotion & social interaction. Fundamentals of Nonverbal Behavior*, pp. 401–432, 1991.

[239] W. S. Condon and W. D. Ogston, "A segmentation of behavior," *Journal of psychiatric research*, vol. 5, no. 3, pp. 221–235, 1967.

[240] J. Vargas-Quiros, Ö. Kapcak, H. Hung, and L. Cabrera-Quiros, "Individual and joint body movement assessed by wearable sensing as a predictor of attraction in speed dates," *IEEE Tran. on Affective Comp.*, vol. 14, no. 3, pp. 2168–2181, 2023.

[241] G. Varni, M. Avril, A. Usta, and M. Chetouani, "Syncpy: A unified open-source analytic library for synchrony," in *Proceedings of the 1st Workshop on Modeling Interpersonal Synchrony and Influence*, 2015, pp. 41–47.

[242] D. Hudson, T. J. Wiltshire, and M. Atzmueller, "Multisyncpy: A python package for assessing multivariate coordination dynamics," *Behavior Research Methods*, vol. 55, no. 2, pp. 932–962, 2023.

[243] S. S. Singh, S. Muhuri, S. Mishra, D. Srivastava, H. K. Shakya, and N. Kumar, "Social network analysis: A survey on process, tools, and application," *ACM Comp. Surveys*, vol. 56, no. 8, 2024.

[244] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Int. Conf. on Learning Representations*, 2018.

[245] Z. Xu, Y. Lin, H. Han, *et al.*, "Mambatalk: Efficient holistic gesture synthesis with selective state space models," *Proc. Neural Inf. Process. Syst.*, vol. 37, pp. 20 055–20 080, 2024.

[246] X. Qi, J. Pan, P. Li, *et al.*, "Weakly-supervised emotion transition learning for diverse 3d co-speech gesture generation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2024, pp. 10 424–10 434.

[247] B. Azari and A. Lim, "Emostyle: One-shot facial expression editing using continuous emotion parameters," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2024, pp. 6385–6394.

[248] Y. Pan, S. Tan, S. Cheng, Q. Lin, Z. Zeng, and K. Mitchell, "Expressive talking avatars," *Transactions on Visualization and Computer Graphics*, vol. 30, no. 5, pp. 2538–2548, 2024.

[249] A. Melnik, M. Miasayedzenkau, D. Makaravets, *et al.*, "Face generation and editing with stylegan: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3557–3576, 2024.

[250] M.-A. Mahfoudi, A. Meyer, T. Gaudin, A. Buendia, and S. Bouakaz, "Emotion expression in human body posture and movement: A survey on intelligible motion factors, quantification and validation," *IEEE Tran. on Affective Computing (TAFFC)*, vol. 14, no. 4, pp. 2697–2721, 2022.

[251] K. Chhatre, N. Athanasiou, G. Becherini, C. Peters, M. J. Black, T. Bolkart, *et al.*, "Emotional speech-driven 3d body animation via disentangled latent diffusion," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2024, pp. 1942–1953.

[252] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, IEEE, vol. 1, 1996, pp. 373–376.

[253] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech synthesis based on hidden markov models," in *Proceedings of the IEEE*, vol. 88, IEEE, 2000, pp. 436–447.

[254] A. van den Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," in *arXiv preprint arXiv:1609.03499*, 2016.

[255] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[256] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 132–157, 2021.

[257] Z. Du, B. Sisman, K. Zhou, and H. Li, "Disentanglement of Emotional Style and Speaker Identity for Expressive Voice Conversion," in *Proc. Interspeech*, Sep. 2022.

[258] F. Kreuk, A. Polyak, J. Copet, *et al.*, "Textless speech emotion conversion using discrete & decomposed representations," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 2022.

[259] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, 2020.

[260] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," in *The Speaker and Language Recognition Workshop (Speaker Odyssey)*, May 2020.

[261] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech synthesis with mixed emotions," *IEEE Tran. on Affective Computing (TAFFC)*, 2022.

[262] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "EmoMix: Emotion Mixing via Diffusion Models for Emotional Speech Synthesis," in *Interspeech*, 2023.

[263] Y. Guo, C. Du, X. Chen, and K. Yu, "Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, 2023.

[264] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Tran. on Affective Computing (TAFFC)*, 2023.

[265] K. Zhou, B. Sisman, and H. Li, "Vaw-gan for disentanglement and recomposition of emotional elements in speech," in *IEEE Spoken Language Tech. Workshop*, 2021.

[266] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021, pp. 920–924.

[267] F. Kreuk, J. Keshet, and Y. Adi, "Self-Supervised Contrastive Learning for Unsupervised Phoneme Segmentation," in *Interspeech*, 2020, pp. 3700–3704.

[268] A. Sicherman and Y. Adi, "Analysing discrete self supervised speech representation for spoken language modeling," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc., ICASSP*, 2023, pp. 1–5.

[269] H.-S. Oh, S.-H. Lee, D.-H. Cho, and S.-W. Lee, "Durflex-evc: Duration-flexible emotional voice conversion leveraging discrete representations without text alignment," *IEEE Tran. on Affective Comp.*, 2025.

[270] A. Polyak, Y. Adi, J. Copet, *et al.*, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Proc. Interspeech*, 2021.

[271] W.-N. Hsu, T. Remez, B. Shi, J. Donley, and Y. Adi, "Revise: Self-supervised speech resynthesis with visual input for universal and generalized speech regeneration," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2023.

[272] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *Journal of Machine Learning Research*, vol. 22, no. 57, pp. 1–64, 2021.

[273] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE Tran. on on Audio, Speech, and Language Processing*, 2023.

[274] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Proc. Neural Inf. Process. Syst.*, vol. 33, pp. 8067–8077, 2020.

[275] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex stft domain," *Proc. Interspeech*, 2022.

[276] B. Lay, S. Welker, J. Richter, and T. Gerkamnn, "Reducing the prior mismatch of stochastic differential equations for diffusion-based speech enhancement," *Interspeech*, 2023.

[277] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *Proc. Int. Conf. Learning Representations*, 2022.

[278] J. Yao, Y. Yang, Y. Lei, *et al.*, "Promptvc: Flexible stylistic voice conversion in latent space driven by natural language prompts," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, IEEE, 2024, pp. 10 571–10 575.

[279] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Int. Conf. Machine Learning (ICML)*, 2016.

[280] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The msp-conversation corpus," *Interspeech*, 2020.

[281] L. I.-K. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, vol. 45, no. 1, p. 255, Mar. 1989.

[282] S. Mohamed and B. Lakshminarayanan, "Learning in implicit generative models," in *Proc. Int. Conf. Learning Representations*, ICLR 2017 Invite to Workshop, 2017.

[283] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," in *Proc. Int. Conf. Learning Representations*, 2017.

[284] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are GANs created equal? a large-scale study," in *Proc. Neural Inf. Process. Syst.*, 2018.

[285] Q. Wu, Y. Liu, H. Zhao, *et al.*, "Uncovering the disentanglement capability in text-to-image diffusion models," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Jun. 2023, pp. 1900–1910.

[286] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, 1982.

[287] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Oct. 2023, pp. 3836–3847.

[288] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd. New York, USA: Springer, 1996.

[289] B. Efron, "Tweedie's formula and selection bias," *Journal of the American Statistical Association*, vol. 106, no. 496, 2011.

[290] F. Wang, Y. Liu, K. Liu, Y. Wang, S. Medya, and P. S. Yu, "Uncertainty in graph neural networks: A survey," *Transactions on Machine Learning Research*, 2024.

[291] E. Hajiramezanali, A. Hasanzadeh, K. Narayanan, N. Duffield, M. Zhou, and X. Qian, "Variational graph recurrent neural networks," *Proc. Neural Inf. Process. Syst.*, vol. 32, 2019.

[292] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *Proc. Neural Inf. Process. Syst. Workshop on Bayesian Deep Learning*, 2016.

[293] Y. Rong, W. Huang, T. Xu, and J. Huang, "Dropedge: Towards deep graph convolutional networks on node classification," in *Proc. Int. Conf. Learning Representations*, 2020.

[294] Y. Zhang, S. Pal, M. Coates, and D. Ustebay, "Bayesian graph convolutional neural networks for semi-supervised classification," in *Proceedings of the AAAI Conf. on Artificial Intelligence*, vol. 33, 2019, pp. 5829–5836.

[295] I. Higgins, L. Matthey, A. Pal, *et al.*, "Beta-vae: Learning basic visual concepts with a constrained variational framework," *Proc. Int. Conf. Learning Representations*, 2017.

[296] H. Kim and A. Mnih, "Disentangling by factorising," in *Proc. Int. Conf. Machine Learning*, PMLR, 2018, pp. 2649–2658.

[297] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. Int. Conf. Machine Learning*, PMLR, 2020, pp. 9929–9939.

[298] B. Biancardi, S. Dermouche, and C. Pelachaud, "Adaptation mechanisms in human–agent interaction: Effects on user's impressions and engagement," *Frontiers in Computer Science*, vol. 3, p. 696 682, 2021.

List of Acronyms

t-SNE	t-distributed Stochastic Neighbor Embedding
AER	automatic emotion recognition
AI	Artificial Intelligence
ANS	autonomic nervous system
ASR	automatic speech recognition
BBB	Bayes by backpropagation
BNN	Bayesian neural network
CCC	concordance correlation coefficient
CNN	convolutional neural network
CNS	central nervous system
DNN	deep neural network
DP	duration predictor
ELBO	evidence lower bound
ESS	emotion-conditioned speech synthesis
EWE	evaluator-weighted means
FER	facial emotion recognition
GAN	generative adversarial network
GAT	graph attention networks
GCN	graph convolution networks
GER	group emotion recognition
GMM	Gaussian mixture model
GNN	graph neural networks
HiFiGAN	high-fidelity generative adversarial network
KL	Kullback–Leibler divergence
LDL	label distribution learning
LDM	latent diffusion models
LLM	large language model
LSTM	long short-term memory

MC	Monte Carlo
MEMO	Memory Aware Conversational AI
ML	machine learning
MLP	multi-layer perceptron
MSE	mean squared error
MTL	multi-task learning
NES	neuro-endocrine system
PCC	Pearson's correlation coefficient
PER	physiological emotion recognition
ReVISE	Resynthesis with Visual Input for Speech REgeneration
SDE	stochastic differential equation
SDS	speech dialogue system
SEC	speech emotion conversion
SER	speech emotion recognition
SLM	speech language model
SNS	somatic nervous system
SSL	self-supervised learning
SSP	social signal processing
STFT	short-time Fourier transform
STL	single-task learning
SV	speaker verification
SVM	support vector machine
SVT	supralaryngeal vocal tract
TTS	text-to-speech synthesis
VAE	variational auto-encoder
VC	voice conversion
VGAE	variational graph auto-encoders

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

Hamburg, 04.12.2025 

Navin Laxminarayanan Raj Prabhu

Appendices

A

Related Peer-Reviewed Publications

A.1 Bayesian Neural Networks for Label Uncertainty Modeling [P2]

Abstract

Emotions are subjective constructs. Recent end-to-end speech emotion recognition systems are typically agnostic to the subjective nature of emotions, despite their state-of-the-art performance. In this work, we introduce an end-to-end Bayesian neural network architecture to capture the inherent subjectivity in the arousal dimension of emotional expressions. To the best of our knowledge, this work is the first to use Bayesian neural networks for speech emotion recognition. At training, the network learns a distribution of weights to capture the inherent uncertainty related to subjective arousal annotations. To this end, we introduce a loss term that enables the model to be explicitly trained on a distribution of annotations, rather than training them exclusively on mean or gold-standard labels. We evaluate the proposed approach on the AVEC'16 dataset. Qualitative and quantitative analysis of the results reveals that the proposed model can aptly capture the distribution of subjective arousal annotations, with state-of-the-art results in mean and standard deviation estimations for uncertainty modeling.

Reference

N. Raj Prabhu and G. Carbajal and N. Lehmann-Willenbrock and T. Gerkmann, *End-To-End Label Uncertainty Modeling for Speech-based Arousal Recognition Using Bayesian Neural Networks*, 151-155, 2022. DOI: 10.21437/Interspeech.2022-10490

Copyright Notice

Authors' Contributions

Navin Raj Prabhu led the study, including the initial conceptualization, algorithm development, neural network training, experimental validation, and manuscript preparation. Guillaume Carbajal contributed with offering insights into the experimental validation and also participated in the manuscript review. Nale Lehmann-Willenbrock contributed by reviewing the manuscript and helping to refine the argumentation and overall framing. Timo Gerkmann provided key insights into the experimental validation, offered valuable methodological feedback through discussions, and participated in the manuscript review.

End-To-End Label Uncertainty Modeling for Speech-based Arousal Recognition Using Bayesian Neural Networks

Navin Raj Prabhu[⋆†] *Guillaume Carbajal*[⋆] *Nale Lehmann-Willenbrock*[†] *Timo Gerkmann*[⋆]

[⋆]Signal Processing, Universität Hamburg, Germany
[†]Industrial and Organizational Psychology, Universität Hamburg, Germany
{navin.raj.prabhu, guillaume.carbajal, nale.lehmann-willenbrock, timo.gerkmann}@uni-hamburg.de

Abstract

Emotions are subjective constructs. Recent end-to-end speech emotion recognition systems are typically agnostic to the subjective nature of emotions, despite their state-of-the-art performance. In this work, we introduce an end-to-end Bayesian neural network architecture to capture the inherent subjectivity in the arousal dimension of emotional expressions. To the best of our knowledge, this work is the first to use Bayesian neural networks for speech emotion recognition. At training, the network learns a distribution of weights to capture the inherent uncertainty related to subjective arousal annotations. To this end, we introduce a loss term that enables the model to be explicitly trained on a distribution of annotations, rather than training them exclusively on mean or gold-standard labels. We evaluate the proposed approach on the AVEC'16 dataset. Qualitative and quantitative analysis of the results reveals that the proposed model can aptly capture the distribution of subjective arousal annotations, with state-of-the-art results in mean and standard deviation estimations for uncertainty modeling.

Index Terms: Bayesian networks, end-to-end speech emotion recognition, uncertainty, subjectivity, label distribution learning

1. Introduction

While individual subjective emotional experiences may be accessed using self-report surveys [1], expressions of emotions are embedded in a social context, which makes them inherently dynamic and subjective interpersonal phenomena [2, 3]. One way in which emotions become expressed in social interactions and therefore accessible for social signal processing concerns speech signals. Speech emotion recognition (SER) research spans roughly two decades [4], with ever improving state-of-the-art results. As a consequence, affective sciences and SER has shown increasing prominence in high-critical and socially relevant domains, e.g. health and employee well-being [4, 5].

Common SER approaches rely on hand-crafted features to predict gold-standard emotion labels [6, 7]. Recently, end-to-end deep neural networks (DNNs) have been shown to deliver state-of-the-art emotion predictions [8, 9], by *learning* features rather than relying on hand-crafted features. Despite their state-of-the-art results, they are often agnostic to the subjective nature of emotions and the resulting label uncertainty [10, 7], thereby inducing limited reliability in SER [4]. However, for any real-world application context, it is crucial that SER systems should not only deliver mean or gold-standard predictions but also account for subjectivity based confidence measures [11, 4].

Han et al., [10, 6] pioneered uncertainty modeling in SER using a multi-task framework to also predict the standard deviation of emotion annotations. Sridhar et al., [7] introduced a dropout-based model to estimate uncertainties. However, these uncertainty models were not trained on the distribution of emotion annotations and relied on hand-crafted features. Of note, in contrast to hand-crafted features, learning representations in an end-to-end manner implies that they are also dependent on the level of subjectivity in label annotations [12].

In machine learning, two types of uncertainty can be distinguished. *Aleatoric* uncertainty captures data inherent noise (label uncertainty) whereas *epistemic* uncertainty accounts for the model parameters and structure (model uncertainty) [13]. Stochastic and probabilistic models have mainly been deployed for uncertainty modeling, using auto-encoder architectures [14], neural processes [15], and Bayesian neural networks (BNN) [16]. Bayes by Backpropagation (BBB) for BNNs [16] uses simple gradient updates to optimize weight distributions. Further with its capability to produce stochastic outputs, it is a promising candidate for end-to-end uncertainty SER.

As opposed to emotions as inner subjective experiences [1], we focus on emotional expressions as behaviors that others subjectively perceive and respond to. A common framework for analyzing the expression of emotion is pleasure-arousal theory [17, 18], which describes emotional experiences in two continuous, bipolar, and orthogonal dimensions: pleasure-displeasure (*valence*) and activation-deactivation (*arousal*). It is documented in SER literature that the audio modality insufficiently explains valence [8, 9]. Noting this, in this work, we decided to specifically focus on the label uncertainty in the *arousal* dimension of emotional expressions.

In this paper, we propose an end-to-end BBB-based BNN architecture for SER. To the best of our knowledge, this is the first time a BNN is used for SER. In contrast to [10, 6, 7], the model can be explicitly trained on a distribution of emotion annotations. For this, we introduce a loss term that promotes capturing *aleotoric* uncertainty (label uncertainty) rather than exclusively capturing *epistemic* uncertainty (model uncertainty). Finally, we show that our proposed model trained on the loss term can aptly capture label uncertainty in arousal annotations.

The rest of the paper is organized as follows. In Section 2, we present related background on label uncertainty. In Section 3, we introduce the proposed end-to-end BNN SER model. In Section 4, we explain the experimental setup. In Section 5, we present the results and raise discussions on them.

2. On uncertainty in arousal annotations

2.1. Ground-truth labels

A crucial challenge in studying emotions within the arousal and valence framework concerns the significant degree of subjectivity surrounding them [4, 10, 19]. To tackle this, annotations

$\{y_1, y_2, .., y_a\}$ for emotions are collected from a annotators [20, 21]. The *ground-truth label* is then obtained as the mean m over all annotations from a annotators [22, 23],

$$m = \frac{1}{a}\sum_{i=1}^{a} y_i. \quad (1)$$

Alternatively, an evaluator-weighted mean has been proposed and referred to as the *gold-standard* $\widetilde{m}$ [24, 8]. However, in order to better represent subjective annotations, in this paper we use the mean m rather than the evaluator-weighted mean.

2.2. Point estimates in SER

Given a raw audio sequence of T frames $\mathcal{X} = [x_1, x_2, ..., x_T]$, typically, the goal is to estimate the ground-truth label m_t for each time frame $t \in [1, T]$, referred to as $\widehat{m}_t$. The concordance correlation coefficient (CCC), which takes both linear correlations and biases into consideration, has been widely used as a loss function for this task. For Pearson correlation r, the CCC between the ground-truth label m and its estimate $\widehat{m}$, for T frames, is formulated as

$$\mathcal{L}_{\text{CCC}}(m) = \frac{2r\sigma_m\sigma_{\widehat{m}}}{\sigma_m^2 + \sigma_{\widehat{m}}^2 + (\mu_m - \mu_{\widehat{m}})^2}, \quad (2)$$

where $\mu_m = \frac{1}{T}\sum_{t=1}^{T} m_t$, $\sigma_m^2 = \frac{1}{T}\sum_{t=1}^{T}(m_t - \mu_m)^2$, and $\mu_{\widehat{m}}, \sigma_{\widehat{m}}^2$ are obtained similarly for $\widehat{m}$.

Early approaches relied on hand-crafted features as inputs to estimate CCC. End-to-end DNNs which circumvent the limitations of hand-crafted and -chosen features have been deployed to yield state-of-the-art performance [25, 8, 9]. Notwithstanding their performance, end-to-end DNNs are trained exclusively on m and therefore cannot account for annotator subjectivity.

2.3. Uncertainty modeling in SER

Han et al. quantified label uncertainty using a statistical estimate, the perception uncertainty s defined as the unbiased standard deviation of a annotators [10, 6]:

$$s = \sqrt{\frac{1}{a-1}\sum_{i=1}^{a}(y_i - m)^2}. \quad (3)$$

They proposed a multi-task model to additionally estimate s along with m. However, the model only accounts for the standard deviation of a annotations, rather than the whole distribution in itself. Thereby, susceptible to unreliable s estimates for lower values of a and sparsely distributed annotations.

Sridhar et al. introduced a Monte-Carlo dropout-based uncertainty model, to obtain stochastic predictions and uncertainty estimates [7]. The model is trained exclusively on m rather than the distribution of annotations, thereby only capturing the model uncertainty and not the label uncertainties.

3. End-to-end label uncertainty model

In order to better represent subjectivity in emotional expressions, we propose to estimate the *emotion annotation distribution* $\mathcal{Y}_t$ for each frame t. While the true distributional family of $\mathcal{Y}_t$ is unknown, we assume, for simplicity, that it follows a Gaussian distribution:

$$\mathcal{Y}_t \sim \mathcal{N}(m_t, s_t). \quad (4)$$

Thus, the goal is to obtain an estimate $\widehat{\mathcal{Y}}_t$ of $\mathcal{Y}_t$ and infer both $\widehat{m}_t$ and $\widehat{s}_t$ from realizations of $\widehat{\mathcal{Y}}_t$.

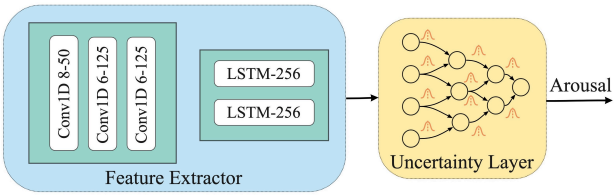


Figure 1: *The proposed architecture. Blue denotes layers with deterministic weights and yellow with probabilistic weights.*

3.1. End-to-end DNN architecture

We propose an end-to-end architecture which uses a feature extractor to learn temporal-paralinguistic features from x_t, and an uncertainty layer to estimate $\mathcal{Y}_t$ (see Fig. 1). The feature extractor, inspired from [8], consists of three Conv1D layers followed by two stacked long-short term memory (LSTM) layers. The uncertainty layer is devised using the BBB technique [16], comprising three BBB-based MLP.

3.2. Model uncertainty loss

Unlike a standard neuron which optimizes a deterministic weight w, the BBB-based neuron learns a probability distribution on the weight w by calculating the variational posterior $P(w|\mathcal{D})$ given the training data $\mathcal{D}$ [16]. Intuitively, this regularizes w to also capture the inherent uncertainty in $\mathcal{D}$.

Following [16], we parametrize $P(w|\mathcal{D})$ using a Gaussian $\mathcal{N}(\mu_w, \sigma_w)$. For non-negative σ_w, we re-parametrize the standard deviation $\sigma_w = \log(1 + \exp(\rho_w))$. Then, $\theta = (\mu_w, \rho_w)$ can be optimized using simple backpropagation.

For an optimized θ, the predictive distribution $\widehat{\mathcal{Y}}_t$ for an audio frame x_t, is given by $P(\widehat{y}_t|x_t) = \mathbb{E}_{P(w|D)}[P(\widehat{y}_t|x_t, w)]$, where $\widehat{y}_t$ are realizations of $\widehat{\mathcal{Y}}_t$. Unfortunately, the expectation under the posterior of weights is intractable. To tackle this, the authors in [16] proposed to learn θ of a weight distribution $q(w|\theta)$, the variational posterior, that minimizes the Kullback-Leibler (KL) divergence with the true Bayesian posterior, resulting in the negative evidence lower bound (ELBO),

$$f(w, \theta)_{\text{BBB}} = \text{KL}\big[q(w|\theta)\|P(w)\big] - \mathbb{E}_{q(w|\theta)}\big[\log P(D|w)\big]. \quad (5)$$

Stochastic outputs in BBB are achieved using multiple forward passes n with stochastically sampled weights w, thereby modeling $\widehat{\mathcal{Y}}_t$ using the n stochastic estimates. To account for the stochastic outputs, (5) is approximated as,

$$\mathcal{L}_{\text{BBB}} \approx \sum_{i=1}^{n} \log q(w^{(i)}|\theta) - \log P(w^{(i)}) - \log P(D|w^{(i)}). \quad (6)$$

where $w^{(i)}$ denotes the i^{th} weight sample drawn from $q(w|\theta)$. The BBB window-size b controls how often new weights are sampled for time-continuous SER. The degree of uncertainty is assumed to be constant within this time period.

During testing, the uncertainty estimate $\widehat{s}_t$ is the standard deviation of $\widehat{\mathcal{Y}}_t$. Similarly, $\widehat{m}_t$ is the realization $\widehat{y}_t$ obtained using the mean of the optimized weights μ_w. Obtaining $\widehat{m}_t$ using μ_w helps overcome the randomization effect of sampling from $q(w|\theta)$, which showed better performances in our case.

3.3. Label uncertainty loss

Inspired by [13], we introduce a KL divergence-based loss term as a measurement of distribution similarity to explicitly fit our model to the label distribution $\mathcal{Y}_t$:

$$\mathcal{L}_{\text{KL}} = f(\mathcal{Y}_t \| \widehat{\mathcal{Y}}_t)_{\text{KL}} = \int \mathcal{Y}_t(x) \, \log \frac{\mathcal{Y}_t(x)}{\widehat{\mathcal{Y}}_t(x)} dx. \quad (7)$$

The KL divergence is asymmetric, making the order of distributions crucial. In 7, the true distribution $\mathcal{Y}_t$ is followed by its estimate $\widehat{\mathcal{Y}}_t$, promoting a mean-seeking approximation rather than a mode-seeking one and capturing the full distribution [26].

3.4. End-to-end uncertainty loss

The proposed end-to-end uncertainty loss is formulated as,

$$\mathcal{L} = \mathcal{L}_{\text{CCC}}(m) + \mathcal{L}_{\text{BBB}} + \alpha \mathcal{L}_{\text{KL}}. \quad (8)$$

Intuitively, $\mathcal{L}_{\text{CCC}}(m)$ optimizes for mean predictions m, $\mathcal{L}_{\text{BBB}}$ optimizes for BBB weight distributions, and $\mathcal{L}_{\text{KL}}$ optimizes for the label distribution $\mathcal{Y}_t$. For $\alpha = 0$, the model only captures model uncertainty (*MU*). For $\alpha = 1$, the model also captures *label uncertainty* (*+LU*). $\mathcal{L}_{\text{CCC}}(m)$ is used as part of $\mathcal{L}$ to achieve faster convergence and jointly optimize for mean predictions. Including $\mathcal{L}_{\text{CCC}}(m)$ might lead to better optimization of the feature extractor, as previously illustrated by [8, 27].

4. Experimental Setup

4.1. Dataset

To validate our model, we use the AVEC'16 emotion recognition dataset [23]. Multimodal signals were recorded from 27 subjects, and in this work we only utilize the audio signals collected at 16 kHz. The dataset consists of continuous arousal annotations by $a = 6$ annotators at 40 ms frame-rate, and post-processed with local normalization and mean filtering. The arousal annotations are distributed on average with $\mu_m = 0.01$ and $\mu_s = 0.23$, where $\mu_s = \frac{1}{T}\sum_{t=1}^{T} s_t$. The dataset is divided into speaker disjoint partitions for training, development and testing, with nine 300 s recordings each. As the annotations for the test partition in the dataset are not publicly available, all results are computed on the development partition.

4.2. Baselines

As baselines, we use Han et al.'s perception uncertainty model (*MTL PU*) and single-task learning model (*STL*) [6]. The *MTL PU* model uses a multi-task technique followed by a dynamic tuning layer to account for perception uncertainty s in the final mean estimations m. The *STL* on the other hand only performs a single task of estimating m.

For a fair comparison, we reimplemented the baselines and tested them in our test bed. Crucially, the reimplementation also enables us to compare the models in-terms of their s estimates, which were not presented in Han et al.'s work [6]. The only difference between Han et al.'s test bed and ours is the post-processing pipeline. While Han et al. use the post-processing pipeline suggested in the AVEC'16 [23], here we use the median filtering [8] as the sole post-processing technique to make all considered approaches comparable.

4.3. Choice of hyperparameters

The hyperparameters of the *feature extractor* (e.g. kernel sizes, filters) are adopted from [27]. A similar extractor with the same hyperparameters has been used in several multimodal emotion recognition tasks with state-of-the-art performance [9, 27].

As the *prior distribution* $P(w)$, [16] recommend a mixture of two Gaussians, with zero means and standard deviations as $\sigma_1 > \sigma_2$ and $\sigma_2 \ll 1$, thereby obtaining a spike-and-slab prior with heavy tail and concentration around zero mean. But in our case, we do not need mean centered predictions as $\mathcal{Y}$ does not follow such a distribution, as seen in Section 4.1. In this light, we propose to use a simple Gaussian prior with unit standard deviation $\mathcal{N}(0, 1)$. The μ_w and ρ_w of the *posterior distribution* $P(w|D)$ are initialized uniformly in the range $[-0.1, 0.1]$ and $[-3, -2]$ respectively. The ranges were fine-tuned using grid search for maximized $\mathcal{L}_{\text{KL}}$ at initialization on the train partition.

It is computationally expensive to sample new weights at every time-step (40 ms) and also the level of uncertainties varies rather slowly. In this light, we set *BBB window-size* $b = 2$ s (50 frames). The same window-size is also used for median filtering, the sole post-processing technique used. The *number of forward passes* in (6) is fixed to $n = 30$, with the time-complexity in consideration.

For training, we use the Adam optimizer with a learning rate of 10^{-4}. The batch size used was 5, with a sequence length of 300 frames, 40 ms each and 12 s in total. Dropout with probability 0.5 was used to prevent overfitting. All the models were trained for a fixed 100 epochs. The best model is selected and used for testing when best $\mathcal{L}$ is observed on train partition.

Table 1: *Comparison on mean m, standard deviation s, and label distribution estimations $\mathcal{Y}$, in terms of $\mathcal{L}_{ccc}(m)$, $\mathcal{L}_{ccc}(s)$, and $\mathcal{L}_{KL}$ respectively. Larger CCC indicates improved performance. Lower KL indicates improved performance. ** indicates statistically significant better results over **all** other approaches in comparison, and * over **only some** of the approaches.*

	$\mathcal{L}_{\text{ccc}}(m)$	$\mathcal{L}_{\text{ccc}}(s)$	$\mathcal{L}_{\text{KL}}$
STL [6]	0.719	-	-
MTL PU [6]	0.734	0.286	0.797
MU	**0.756***	0.076	0.690
+LU	0.744	**0.340****	**0.258****

4.4. Validation metrics

To validate our mean estimates, we use $\mathcal{L}_{\text{CCC}}(m)$, a widely used metric in literature [8, 27, 6]. To validate the label uncertainty estimates, we use $\mathcal{L}_{\text{CCC}}(s)$, along with $\mathcal{L}_{\text{KL}}$. $\mathcal{L}_{\text{CCC}}(s)$, previously used in [10], only validates the performance on s, and ignores performances on m. In this light, we additionally use $\mathcal{L}_{\text{KL}}$ which jointly validates both m and s performances. However, the metric makes a Gaussian assumption on $\mathcal{Y}_t$ and hence can be biased. In this light, we use both these metrics for the validation with their respective benefits under consideration. Statistical significance is estimated using one-tailed t-test, asserting significance for p-values ≤ 0.05, similar to [28].

5. Results and Discussion

Table 1 shows the average performance of the baselines and the proposed models, in terms of their mean m, standard deviation s, and distribution $\widehat{\mathcal{Y}}_t$ estimations, $\mathcal{L}_{\text{CCC}}(m)$, $\mathcal{L}_{\text{CCC}}(s)$ and $\mathcal{L}_{KL}$ respectively. Comparisons with respect to $\mathcal{L}_{\text{CCC}}(s)$ and $\mathcal{L}_{KL}$ are not presented for the STL baseline as this algorithm does not contain uncertainty modeling and does not estimate s.

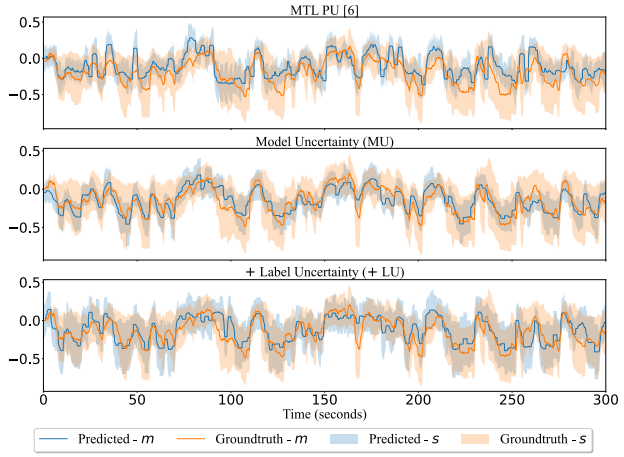


Figure 2: *Results obtained for a test subject for arousal.*

5.1. Comparison with baselines

From Table 1, we observe that both the proposed uncertainty models, MU and +LU, provide improved mean m estimations over the baselines, in-terms of $\mathcal{L}_{\text{CCC}}(m)$, with statistical significance. Crucially, we observe that our models provide better $\mathcal{L}_{\text{CCC}}(m)$ even in comparison with the STL baseline [6] which is not an uncertainty model and only estimates m without accounting for the label uncertainty. While uncertainty models MU and +LU achieve 0.756 and 0.744 $\mathcal{L}_{\text{CCC}}(m)$ respectively, STL and MTL PU [6] achieve 0.734 and 0.719 respectively.

Secondly, we note that the proposed label uncertainty model +LU achieves state-of-the-art results for standard deviation s and distribution $\widehat{\mathcal{Y}}_t$ estimations, with statistical significance, achieving 0.340 $\mathcal{L}_{\text{CCC}}(s)$ and 0.258 $\mathcal{L}_{KL}$ respectively. The +LU model is trained on a more informative distribution of annotations $\mathcal{Y}_t$, in contrast to training on the s estimate [6], thereby leading to better capturing label uncertainty in arousal annotations. This explains the capability of distribution learning, using the $\mathcal{L}_{\text{KL}}$ loss term, for modeling label uncertainty in SER. Also noting here that, in contrast to the baselines, the proposed model learns uncertainty dependent representations in an end-to-end manner for improved uncertainty estimates, inline with literature [12] which recommends end-to-end learning for uncertainty modeling in SER. Conclusively, the results in-terms of $\mathcal{L}_{\text{CCC}}(m)$, $\mathcal{L}_{\text{CCC}}(s)$, and $\mathcal{L}_{KL}$ reveal the ability of the proposed +LU model to best capture label uncertainty, thereby also improving mean estimations in comparison to the baselines.

To further validate the results, we plot the distribution estimations for a test subject, seen in Figure 2. From the figure, as the quantitative results suggest, we see that the MU and +LU models are superior to the baseline MTL PU in-terms of the distribution estimations. Specifically, the +LU captures the whole distribution of arousal annotations better than the MTL PU model, by well capturing the time-varying uncertainty without noisy estimates. This further highlights the advantage of training on $\mathcal{L}_{\text{KL}}$ loss term.

5.2. Comparison between *MU* and *+LU*

Between the proposed uncertainty models, we note that +LU outperforms MU in terms of $\mathcal{L}_{\text{CCC}}(s)$, without significant degradations to $\mathcal{L}_{\text{CCC}}(m)$. Specifically, +LU improves drastically and significantly in terms of $\mathcal{L}_{\text{CCC}}(s)$ from 0.076 to 0.340. At the same time, for $\mathcal{L}_{\text{CCC}}(m)$ a slight reduction from 0.756 to 0.744 can be observed which, however, is not statistically significant. This reveals that, by also optimizing $\mathcal{L}_{\text{KL}}$, our model can better account for subjectivity in arousal. It is important to note here that the trade-off between $\mathcal{L}_{\text{CCC}}(s)$ and $\mathcal{L}_{\text{CCC}}(s)$ can be further adjusted using a different prior $P(w)$ on the weight distributions, as noted by Blundell et al. [16] who recommend a spike-and-slab for mean centered predictions. However, in our case, as the arousal annotations do not follow a mean centered distribution (seen in Section 4.1), we chose a simple Gaussian prior with unit standard deviation $\mathcal{N}(0, 1)$ to better capture the annotation distribution. Moreover, the choice of such as simple prior makes our model also scalable, eliminating the requirement to tune the prior with respect to different SER datasets.

Moreover, from the Figure 2, backing the results in Table 1, one may see that +LU, explicitly trained on $\mathcal{L}_{\text{KL}}$, best captures the subjectivity in emotions, while MU is optimized more for predictions centered on the mean and strict standard deviations.

5.3. Label Uncertainty BNN for SER

This work is the first in literature to study BNNs for SER. The BBB technique, adopted by the proposed model, use simple gradient updates and produce stochastic outputs, making them promising candidates for end-to-end uncertainty modeling. Crucially, they open up possibilities for training the model on a distribution of annotations, rather than a less informative standard deviation estimate. Moreover, unlike the MTL PU which requires dataset-dependent tuning of the loss function, using the correlation estimate between m and s [6], our proposed models do not require loss function tuning and are scalable across SER datasets with a simple prior initialization.

While the proposed model has several advantages, it also leaves room for future work. Firstly, the heuristics-based initialization of $P(w|D)$ for optimal performances could be further studied. Secondly, in the model we assumed that $\mathcal{Y}_t$ follows a Gaussian distribution, and had only $a = 6$ annotations available to model the distribution. This might sometimes lead to unstable $\mathcal{L}_{\text{KL}}$ during approximation of $\mathcal{Y}_t$, thereby affecting the training processes. Acknowledging that gaining more annotation is resource inefficient, as future work, we will investigate techniques to model stable $\mathcal{Y}_t$ with limited annotations.

6. Conclusions

We introduced a BNN-based end-to-end approach for SER, which can account for the subjectivity and label uncertainty in emotional expressions. To this end, we introduced a loss term based on the KL divergence to enable our approach to be trained on a distribution of annotations. Unlike previous approaches, the stochastic outputs of our approach can be employed to estimate statistical moments such as the mean and standard deviation of the emotion annotations. Analysis of the results reveals that the proposed uncertainty model trained on the KL loss term can aptly capture the distribution of arousal annotations, achieving state-of-the-art results in mean and standard deviation estimations, in-terms of both the CCC and KL divergence metrics.

7. Acknowledgements

The authors thank the Landesforschungsförderung Hamburg (LFF-FV79) for supporting this work under the "Mechanisms of Change in Dynamic Social Interaction" project.

8. References

[1] J. E. LeDoux and S. G. Hofmann, "The subjective experience of emotion: a fearful view," *Current Opinion in Behavioral Sciences*, vol. 19, pp. 67–72, 2018.

[2] Z. Lei and N. Lehmann-Willenbrock, "Affect in meetings: An interpersonal construct in dynamic interaction processes," in *The Cambridge handbook of meeting science*. Cambridge University Press, 2015, pp. 456–482.

[3] L. Nummenmaa, R. Hari, J. K. Hietanen, and E. Glerean, "Maps of subjective feelings," *Proceedings of the National Academy of Sciences*, vol. 115, no. 37, pp. 9198–9203, 2018.

[4] B. W. Schuller, "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.

[5] D. Dukes, K. Abrams, R. Adolphs, M. E. Ahmed, A. Beatty, K. C. Berridge, S. Broomhall, T. Brosch, J. J. Campos, Z. Clay *et al.*, "The rise of affectivism," *Nature Human Behaviour*, pp. 1–5, 2021.

[6] J. Han, Z. Zhang, Z. Ren, and B. Schuller, "Exploring perception uncertainty for emotion recognition in dyadic conversation and music listening," *Cognitive Computation*, pp. 1–10, 2020.

[7] K. Sridhar and C. Busso, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *ICASSP - IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2020, pp. 8384–8388.

[8] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," in *ICASSP - Int. Conf. on Acoustics, Speech and Signal Processing*, 2018, pp. 5089–5093.

[9] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Information Fusion*, vol. 68, pp. 46–53, 2021.

[10] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proceedings of the 25th ACM Int. Conf. on Multimedia*, 2017, pp. 890–897.

[11] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[12] S. Alisamir and F. Ringeval, "On the evolution of speech representations for affective computing: A brief history and critical overview," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 12–21, 2021.

[13] R. Zheng, S. Zhang, L. Liu, Y. Luo, and M. Sun, "Uncertainty in bayesian deep label distribution learning," *Applied Soft Computing*, 2021.

[14] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, and O. Ronneberger, "A probabilistic u-net for segmentation of ambiguous images," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[15] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. Rezende, and S. A. Eslami, "Conditional neural processes," in *Int. Conf. on Machine Learning*. PMLR, 2018, pp. 1704–1713.

[16] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1613–1622.

[17] R. Reisenzein, "Pleasure-arousal theory and the intensity of emotions." *Journal of personality and social psychology*, vol. 67, no. 3, p. 525, 1994.

[18] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[19] M. K. T, E. Sanchez, G. Tzimiropoulos, T. Giesbrecht, and M. Valstar, "Stochastic Process Regression for Cross-Cultural Speech Emotion Recognition," in *Proc. Interspeech 2021*, 2021, pp. 3390–3394.

[20] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.

[21] N. Raj Prabhu, C. Raman, and H. Hung, "Defining and Quantifying Conversation Quality in Spontaneous Interactions," in *Comp. Pub. of 2020 Int. Conf. on Multimodal Interaction*, 2020, pp. 196–205.

[22] M. Abdelwahab and C. Busso, "Active learning for speech emotion recognition using deep neural network," in *2019 8th Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019.

[23] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 3–10. [Online]. Available: https://doi.org/10.1145/2988257.2988258

[24] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* IEEE, 2005, pp. 381–385.

[25] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 Int. Conf. on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016.

[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, 3rd ed., ser. 5. The address of the publisher: MIT Press, 7 2016, vol. 4, ch. 3, pp. 51–77, http://www.deeplearningbook.org.

[27] P. Tzirakis, A. Nguyen, S. Zafeiriou, and B. W. Schuller, "Speech emotion recognition using semantic information," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6279–6283.

[28] K. Sridhar, W.-C. Lin, and C. Busso, "Generative approach using soft-labels to learn uncertainty in predicting emotional attributes," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

A.2 Bayesian Neural Networks for Label Uncertainty Modeling [P3]

Abstract

As different people perceive others' emotional expressions differently, their annotation in terms of arousal and valence are per se subjective. To address this, these emotion annotations are typically collected by multiple annotators and averaged across annotators in order to obtain labels for arousal and valence. However, besides the average, also the uncertainty of a label is of interest, and should also be modeled and predicted for automatic emotion recognition. In the literature, for simplicity, label uncertainty modeling is commonly approached with a Gaussian assumption on the collected annotations. However, as the number of annotators is typically rather small due to resource constraints, we argue that the Gaussian approach is a rather crude assumption. In contrast, in this work we propose to model the label distribution using a Student's t-distribution which allows us to account for the number of annotations available. With this model, we derive the corresponding Kullback-Leibler divergence based loss function and use it to train an estimator for the distribution of emotion labels, from which the mean and uncertainty can be inferred. Through qualitative and quantitative analysis, we show the benefits of the t-distribution over a Gaussian distribution. We validate our proposed method on the AVEC'16 dataset. Results reveal that our t-distribution based approach improves over the Gaussian approach with state-of-the-art uncertainty modeling results in speech-based emotion recognition, along with an optimal and even faster convergence.

Reference

N. Raj Prabhu and N. Lehmann-Willenbrock and T. Gerkmann, *Label Uncertainty Modeling and Prediction for Speech Emotion Recognition using t-Distributions*, Nara, Japan, 2022. DOI: 10.1109/ACII55700.2022.9953816

Copyright Notice

Authors' Contributions

Navin Raj Prabhu led the study, including the initial conceptualization, mathematical derivations for the loss function, algorithm development, neural network training, experimental validation, and manuscript preparation. Nale Lehmann-Willenbrock contributed by reviewing the manuscript and helping to refine the argumentation and overall framing. Timo Gerkmann provided key insights into the experimental validation and mathematical derivations, offered valuable methodological feedback through discussions, and participated in the manuscript review.

Label Uncertainty Modeling and Prediction for Speech Emotion Recognition using *t*-Distributions

Navin Raj Prabhu
Signal Processing
Universität Hamburg
Hamburg, Germany
navin.raj.prabhu@uni-hamburg.de

Nale Lehmann-Willenbrock
Industrial and Organizational Psychology
Universität Hamburg
Hamburg, Germany
nale.lehmann-willenbrock@uni-hamburg.de

Timo Gerkmann
Signal Processing
Universität Hamburg
Hamburg, Germany
timo.gerkmann@uni-hamburg.de

***Abstract*—As different people perceive others' emotional expressions differently, their annotation in terms of arousal and valence are per se subjective. To address this, these emotion annotations are typically collected by multiple annotators and averaged across annotators in order to obtain labels for arousal and valence. However, besides the average, also the uncertainty of a label is of interest, and should also be modeled and predicted for automatic emotion recognition. In the literature, for simplicity, label uncertainty modeling is commonly approached with a Gaussian assumption on the collected annotations. However, as the number of annotators is typically rather small due to resource constraints, we argue that the Gaussian approach is a rather crude assumption. In contrast, in this work we propose to model the label distribution using a Student's t-distribution which allows us to account for the number of annotations available. With this model, we derive the corresponding Kullback-Leibler divergence based loss function and use it to train an estimator for the distribution of emotion labels, from which the mean and uncertainty can be inferred. Through qualitative and quantitative analysis, we show the benefits of the t-distribution over a Gaussian distribution. We validate our proposed method on the AVEC'16 dataset. Results reveal that our t-distribution based approach improves over the Gaussian approach with state-of-the-art uncertainty modeling results in speech-based emotion recognition, along with an optimal and even faster convergence.**

***Index Terms*—uncertainty, subjectivity, distribution learning, t-distribution, Bayesian networks, speech emotion recognition**

I. Introduction

Emotions can be inner subjective experiences, but in order to become socially relevant, they need to be expressed in social context (e.g., [1]). Therefore, emotions are typically studied as emotional expressions that others subjectively perceive and respond to [2]. A common theoretical backdrop for analyzing emotions is the two-dimensional pleasure and arousal framework [3], which describes emotional expressions in two continuous, bipolar, and orthogonal dimensions: pleasure-displeasure (*valence*) and activation-deactivation (*arousal*). One way in which emotions become expressed in social interactions, and therefore accessible for social signal processing (SSP), concerns speech signals. Speech emotion recognition (SER) research spans roughly two decades [2], with ever improving state-of-the-art results. As a consequence, affective sciences and SER has shown increasing prominence in high-critical and socially relevant domains, e.g. health, security, and employee well-being [2], [4], [5].

This work was supported by the Landesforschungsförderung Hamburg (LFF-FV79), under the "Mechanisms of Change in Dynamic Social Interaction" project.

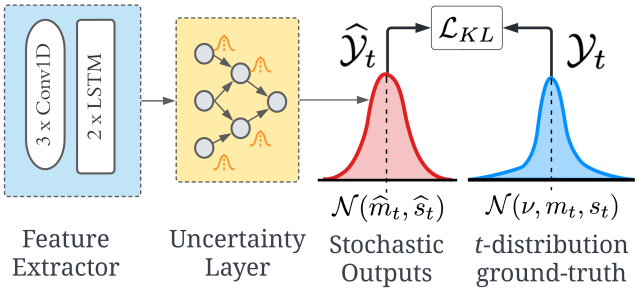


Fig. 1: Overview of the proposed architecture and loss $\mathcal{L}_{\text{KL}}$.

A crucial challenge when studying emotional expressions and trying to establish a ground truth using the pleasure-arousal framework concerns the significant degree of subjectivity surrounding the perceptions of these expressions [2]. Commonly, majority voting [6] or evaluator-weighted mean (EWE) [7] have been used as approximations to obtain ground-truth labels. However, in the context of reliable real-world applications, it is required for SER systems to not only model ground-truth labels but also account for subjectivity based label uncertainty [2], [8].

In SER, label uncertainty has been approached using soft-labels [5], multi-task learning (MTL) [9], [10], stochastic models [11], [12], and label distribution learning [13], [14]. The subjective annotations of emotion creates a label distribution which explains the uncertainty in emotions [5]. In this light, label distribution learning techniques for label uncertainty in SER are gaining research focus, with improved performances [13], [14]. However, a problem with label distribution learning in SER is the limited annotations available [5], [14], due to resource inefficient task of gaining more annotations [5].

Emotion annotations as label distributions are usually modeled by making a Gaussian assumption on them for mathematical convenience [13], [14]. However, a Gaussian assumption with limited samples is not well justified [15], [16], as the central limit theorem (CLT) which primarily backs Gaussian distributions does not hold with insufficient numbers of sam-

ples [17]. Publicly available SER datasets commonly comprise of only three to six annotations [18]–[22], and well agree that gaining more annotations is resource inefficient [5], [23]. In this light, it is important for machine learning (ML) models to account for the limited annotations and model the label distribution accordingly. Alternately, Student's t-distribution, also known as t-distribution, is a probability distribution that also accounts for the number of samples available while modeling [15]. Noting this, and the resource constraints in gaining more annotations, the t-distribution becomes a more appropriate choice for modeling emotion annotations.

In machine learning, two types of uncertainty can be distinguished. *Label uncertainty* captures data inherent noise whereas *model uncertainty* accounts for the uncertainty in model parameters [24], [25]. Stochastic and probabilistic models have mainly been deployed for uncertainty modeling [26]–[28]. Bayes by Backpropagation (BBB) for Bayesian neural networks (BNN) [28] uses *simple gradient updates* to optimize weight distributions for *stochastic outputs*, thereby are promising candidates for label distribution learning in SER.

In this paper, we propose to model emotion annotations as a t-distribution, in contrast to a Gaussian assumption [13], [14]. To the best of our knowledge, this is the first time the problem of limited emotion annotations is tackled from an ML perspective, a common challenge in affective computing and SSP [23]. For this, we adopt a BBB-based stochastic uncertainty model, as proposed in [14], to include a t-distribution instead of a Gaussian. To this end, we introduce a Kullback-Leibler (KL) divergence loss for label uncertainty that quantifies distribution similarity between stochastic emotion predictions, modeled as a Gaussian distribution, and *ground-truth emotion annotations*, modeled as a *t-distribution*. Subsequently, we present analyses to reveal the benefits of using t-distribution over a Gaussian. Finally, we show that the BBB-based uncertainty model trained on the proposed t-distribution based KL-divergence loss can aptly capture label uncertainty with state-of-the-art results, along with a robust loss curve.

II. Related Work

A. Ground-truth labels

To handle subjectivity in emotional expressions, annotations $\{y_1, y_2, .., y_a\}$ for emotions are collected from a annotators [20], [22]. The *ground-truth label* is then obtained as the mean m over all annotations from a annotators [29], [30],

$$m = \frac{1}{a}\sum_{i=1}^{a} y_i. \tag{1}$$

Alternatively, the EWE, which weights annotations with inter-annotator correlations, has been proposed and referred to as the *gold-standard* $\widetilde{m}$ [7]. Both m and $\widetilde{m}$ based approximation of ground-truth leads to loss of information on subjectivity [5].

Traditional SER approaches, given a raw audio sequence of T frames $\mathcal{X} = [x_1, x_2, ..., x_T]$, aim to estimate either the m_t or $\widetilde{m}_t$ for each time frame $t \in [1, T]$, referred to as $\widehat{m}_t$. The concordance correlation coefficient (CCC) has been widely used as a loss function for this task [2]. For Pearson correlation r, the CCC between m and $\widehat{m}$, for T frames, is formulated as

$$\mathcal{L}_{\text{CCC}}(m) = \frac{2r\sigma_m\sigma_{\widehat{m}}}{\sigma_m^2 + \sigma_{\widehat{m}}^2 + (\mu_m - \mu_{\widehat{m}})^2}, \tag{2}$$

where $\mu_m = \frac{1}{T}\sum_{t=1}^{T} m_t$, $\sigma_m^2 = \frac{1}{T}\sum_{t=1}^{T}(m_t - \mu_m)^2$, and $\mu_{\widehat{m}}$, $\sigma_{\widehat{m}}^2$ are obtained similarly for $\widehat{m}$.

B. Label uncertainty in SER

Alternative to exclusively modeling m_t or $\widetilde{m}_t$, works have attempted to model ground truth that also explains inter-annotator disagreement, for example by means of soft labels [5] and entropy of disagreement [31]. Fayek et al. [32] and Tarantino et al. [33] proposed to learn soft labels instead of m_t with improved performance. Steidl et al. [31] quantified label uncertainty using the entropy measure, and trained a model to minimize the difference in entropy between model outputs and annotator disagreement. Sridhar et al. [5] proposed an auto-encoder based learning technique to jointly model soft- and hard-labels of emotion annotations, and subsequently estimating label uncertainty as the entropy on soft-labels.

Label uncertainty has also been approached as a prediction task by estimating either the moments of the distribution [9], [10] or the distribution in itself [13], [14]. Han et al. [9], [10] used an MTL approach to model the unbiased standard deviation s of a annotators as an auxiliary task,

$$s = \sqrt{\frac{1}{a-1}\sum_{i=1}^{a}(y_i - m)^2}. \tag{3}$$

Sridhar et al. [11] introduced a Monte-Carlo dropout model to obtain uncertainty estimates from the distribution of stochastic outputs. Foteinopoulou et al. [13] trained a MTL network using a KL divergence loss that models emotion annotations as a uni-variate Gaussian with mean m and unknown variance. Raj Prabhu et al. [14] introduced a stochastic BNN and trained them on Gaussian emotion annotations. Notwithstanding their improved performances, in [13], [14] *Gaussian* emotion annotations are assumed, despite only having limited annotations. Apart from the apparent mathematical incorrectness of this assumption, they are susceptible to unreliable m and s for lower values of a and sparsely distributed annotations [14].

C. On distributions

A Gaussian distribution $\widehat{\mathcal{Y}} \sim \mathcal{N}(\widehat{\mu}, \widehat{\sigma}^2)$ is a continuous probability distribution for a real-valued random variable y, with general form of its probability density function [16]

$$p(y \mid \widehat{\mu}, \widehat{\sigma}) = \frac{1}{\widehat{\sigma}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\widehat{\mu}}{\widehat{\sigma}}\right)^2}. \tag{4}$$

The parameters $\widehat{\mu}$ and $\widehat{\sigma}$ are the mean and standard deviation of the distribution, respectively.

Due to its simplicity and intelligibility, Gaussian distributions are often used to model random variables whose distribution are unknown [13], [14], [28]. Their importance is however backed by the CLT which only holds *as the number*

of observations of the random variable grows [17]. However, due to the resource constraints in collecting annotations, in most human-behaviour research [23] and in SER [18]–[21], we do not have sufficient annotations to assume a Gaussian distribution on them. As this is a common challenge for reliable real-world applications, it is important for SER algorithms to account for *limited* annotations and model annotation distributions accordingly. Kotz and Nadarajah [15], and, Bishop and Nasrabadi [16], note that in scenarios of limited observations and samples the t-distribution becomes more robust and realistic over a Gaussian.

Student's t-distribution is a probability distribution that arises when estimating the moments of a normally distributed population in *situations where the sample size is small* [15], [34], with the probability density function given by [35]–[37],

$$p(y \mid \nu, \mu, c\sigma) = \frac{1}{\mathrm{B}(\frac{1}{2}, \frac{\nu}{2})} \frac{1}{\sqrt{\nu\sigma^2}} \left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}, \quad (5)$$

where ν denotes the degrees of freedom and B(., .) is the Beta function, for Gamma function Γ, formulated as,

$$B(i, j) = \frac{\Gamma(i)\,\Gamma(j)}{\Gamma(i+j)}. \quad (6)$$

The density function (5) resembles the bell shape of a normally distributed variable, except that it has heavier tails, meaning that it better captures values that fall far from its mean [15], [16]. The degree of freedom ν, also known as the normality parameter, controls the normality of the distribution, and is correlated with the σ parameter [15], [16]. The standard deviation σ in (5) is scaled by ν and is formulated as

$$\sigma = \sigma \sqrt{\frac{\nu}{\nu - 2}} \text{ for } \nu > 2. \quad (7)$$

As ν increases, the t-distribution approaches the normal distribution [37].

III. Proposed t-distribution Label Uncertainty Model

To better represent subjectivity in annotations of emotional expressions, we propose to estimate the *emotion annotation distribution* $\mathcal{Y}_t$ for each frame t. For this, in contrast to a Gaussian assumption $\mathcal{Y}_t \sim \mathcal{N}(m_t, s_t)$ [13], [14], we model the annotations as a t-distribution with degrees of freedom ν:

$$\mathcal{Y}_t \sim \mathcal{N}(\nu, m_t, s_t). \quad (8)$$

Thus, the goal is to obtain an estimate $\widehat{\mathcal{Y}}_t$ of $\mathcal{Y}_t$.

A. Model architecture

We adopt an end-to-end architecture, initially proposed by Raj Prabhu et al. [14], which uses a feature extractor [38] to learn temporal-paralinguistic features from x_t, and a BBB-based uncertainty layer [28] to estimate $\mathcal{Y}_t$. We include the t-distribution modeling as part of the architecture, and the architecture proposed here can be seen in Figure 1.

Unlike a standard neuron which optimizes a deterministic weight w, the BBB-based neuron learns a probability distribution on the weight $P(w|\mathcal{D})$, parametrized by $\theta = (\mu_w, \sigma_w)$ using a Gaussian $\mathcal{N}(\mu_w, \sigma_w)$, given the training data $\mathcal{D}$ [28]. For an optimized θ, the predictive distribution $\widehat{\mathcal{Y}}_t$ for an audio frame x_t, is given by $P(\widehat{y}_t|x_t) = \mathbb{E}_{P(w|D)}[P(\widehat{y}_t|x_t, w)]$, where $\widehat{y}_t$ are realizations of $\widehat{\mathcal{Y}}_t$. Stochastic outputs in BBB are achieved using multiple forward passes n with stochastically sampled weights w, thereby modeling $\widehat{\mathcal{Y}}_t$ using the n stochastic estimates. Following [28], the BBB-based MLP is trained on the negative evidence lower bound (ELBO),

$$\mathcal{L}_{\text{BBB}} \approx \sum_{i=1}^{n} \log q(w^{(i)}|\theta) - \log P(w^{(i)}) - \log P(D|w^{(i)}), \quad (9)$$

where $q(w|\theta)$ is the variational posterior that minimizes the KL divergence with the true Bayesian posterior, and $w^{(i)}$ is the i^{th} sampled weight from $q(w|\theta)$. Finally, as suggested in [14], during testing the uncertainty estimate $\widehat{s}_t$ is the standard deviation of $\widehat{\mathcal{Y}}_t$, and mean estimate $\widehat{m}_t$ is the realization $\widehat{y}_t$ obtained using the mean of the optimized weights μ_w.

B. t-distribution label uncertainty loss derivation

To capture the label uncertainty, we derive a KL divergence based loss function, between the Gaussian stochastic outputs $\widehat{\mathcal{Y}}$ and the t-distribution ground-truth $\mathcal{Y}$. Note that assuming a Gaussian distribution on the stochastic outputs $\widehat{\mathcal{Y}}$ is a fair assumption as the number of stochastic outputs to model $\widehat{\mathcal{Y}}$ can be controlled using n in (9). As the number of sample observations for a distribution approaches thirty a t-distribution converges to a stable Gaussian [36], [37]. Noting this, we intend to choose a n greater than 30 and thereby assume $\widehat{\mathcal{Y}}$ to be Gaussian. As a positive side effect, we result in deriving the KL divergence between a Gaussian and a t-distribution, in contrast to between two t-distributions, with the later involving mathematical complexities in calculating intractable expectations for a loss function.

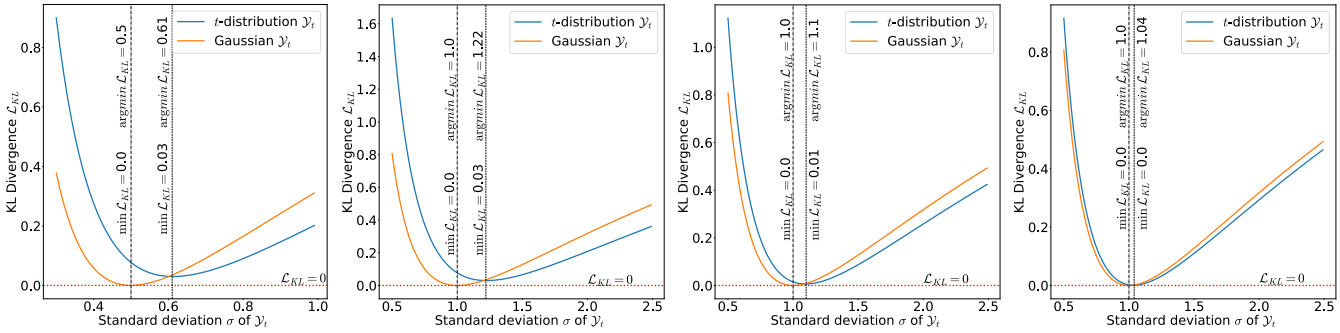
For a Gaussian $\widehat{\mathcal{Y}}$ (see (4)), and a t-distributed $\mathcal{Y}$ (see (5)), the $\mathcal{L}_{KL}$ is formulated as [39], [40],

$$\mathcal{L}_{KL} = f_{KL}(\mathcal{Y}_t \| \widehat{\mathcal{Y}}_t) = H(\mathcal{Y}_t, \widehat{\mathcal{Y}}_t) - H(\mathcal{Y}_t), \quad (10)$$

where H(.,.) is the cross-entropy between two distributions, and H(.) is the entropy of a distribution. Similar to [14], in (10), we choose the true distribution $\mathcal{Y}_t$ to precede its estimate $\widehat{\mathcal{Y}}_t$, promoting a mean-seeking approximation rather than a mode-seeking one and capturing the full distribution [41].

The cross-entropy term H(.,.) in (10), using (4), can be further formulated as,

$$\begin{aligned}
\mathrm{H}(\mathcal{Y}_t, \widehat{\mathcal{Y}}_t) &= -\int \mathcal{Y}_t(y)\ \log\, \widehat{\mathcal{Y}}_t(y)\, dy \\
&= -\int \mathcal{Y}_t(y)\ \Big[\log\Big(\frac{1}{\sqrt{2\pi\widehat{\sigma}^2}} e^{-\frac{1}{2}\left(\frac{y-\widehat{\mu}}{\widehat{\sigma}}\right)^2}\Big)\Big] dy \\
&= -\int \mathcal{Y}_t(y)\ \Big[-\frac{1}{2}\log(2\pi\widehat{\sigma}^2) + \log\Big(e^{-\frac{1}{2}\left(\frac{y-\widehat{\mu}}{\widehat{\sigma}}\right)^2}\Big)\Big] dy \\
&= \frac{1}{2}\log(2\pi\widehat{\sigma}^2) + \int \mathcal{Y}_t(y)\ \Big(\frac{(y-\widehat{\mu})^2}{2\widehat{\sigma}^2}\Big) dy \\
&= \frac{1}{2}\log(2\pi\widehat{\sigma}^2) + \frac{1}{2\widehat{\sigma}^2}\Big[\int \mathcal{Y}_t(y) y^2\, dy - 2\widehat{\mu}\int \mathcal{Y}_t(y)\, y\, dy \\
&\qquad + \widehat{\mu}^2 \int \mathcal{Y}_t(y)\, dy\Big] \quad (11)
\end{aligned}$$



(a) $\widehat{\mathcal{Y}}_t \sim \mathcal{N}(0, 0.5)$, and $\nu = 6$ (b) $\widehat{\mathcal{Y}}_t \sim \mathcal{N}(0, 1)$, and $\nu = 6$ (c) $\widehat{\mathcal{Y}}_t \sim \mathcal{N}(0, 1)$, and $\nu = 12$ (d) $\widehat{\mathcal{Y}}_t \sim \mathcal{N}(0, 1)$, and $\nu = 30$

Fig. 2: Analysis of the t-distribution based KL divergence $\mathcal{L}_{KL}$ (13), in comparison with Gaussian $\mathcal{L}_{KL}$ (14).

Noting that $\int \mathcal{Y}_t(y) y^2 \, dy = \mu^2 + \sigma^2$, $\int \mathcal{Y}_t(y) \, y \, dy = \mu$, and $\int \mathcal{Y}_t(y) \, dy = 1$, where μ and σ are parameters of the t-distribution $\mathcal{Y}_t$, $p(y \mid \nu, \mu, \sigma)$, the equation (11) becomes,

$$\begin{aligned} &= \frac{1}{2}\log(2\pi\widehat{\sigma}^2) + \frac{1}{2\widehat{\sigma}^2}\Big[\sigma^2 + \mu^2 - 2\widehat{\mu}\mu + \widehat{\mu}^2\Big] \\ &= \frac{1}{2}\log(2\pi\widehat{\sigma}^2) + \frac{\sigma^2 + (\mu - \widehat{\mu})^2}{2\widehat{\sigma}^2} \end{aligned} \tag{12}$$

Finally, using (12) in (10), our proposed KL divergence is

$$\boxed{\mathcal{L}_{KL} = \frac{1}{2}\log(2\pi\widehat{\sigma}^2) + \frac{\sigma^2 + (\mu - \widehat{\mu})^2}{2\widehat{\sigma}^2} - H(\mathcal{Y}_t).} \tag{13}$$

We implement (13) as a custom loss function using the pytorch package [42], by extending the `studentT` sub-package[1].

Similarly, as used in [14], the KL divergence between two Gaussians $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}(\widehat{\mu}, \widehat{\sigma}^2)$ is given by [16], [42]

$$\mathcal{L}_{KL} = \log\left(\frac{\widehat{\sigma}}{\sigma}\right) + \frac{\sigma^2 + (\mu - \widehat{\mu})^2}{2\widehat{\sigma}^2} - \frac{1}{2}. \tag{14}$$

While the two loss-functions (14) and (13) have their second term in common, two differences can be noted. Firstly, as (14) calculates the divergence between two similar distributions, $\mathcal{Y}_t$ and $\widehat{\mathcal{Y}}_t$, (14) includes the logarithm of the ratio between the two Gaussian's standard deviation in its formulation. However, in (13), the deviations of $\widehat{\mathcal{Y}}_t$ and $\mathcal{Y}_t$ are separately quantified using terms $\frac{1}{2}\log(2\pi\widehat{\sigma}^2)$ and $H(\mathcal{Y}_t)$, respectively. Secondly, the number of annotators is included in (13) by scaling σ_t with normality factor ν, using (7). The implication of these differences, and the quantitative differences between (14) and (13) are presented and analyzed in the following section.

C. t-distribution loss analysis

In contrast to [14], where the loss function is $\mathcal{L}_{KL}$ between two Gaussians, and to [13], where $\mathcal{L}_{KL}$ is between a Gaussian and a Dirac delta, our proposed loss (13) formulates the KL divergence between a Gaussian and a t-distribution to capture label uncertainty when only limited annotations are available.

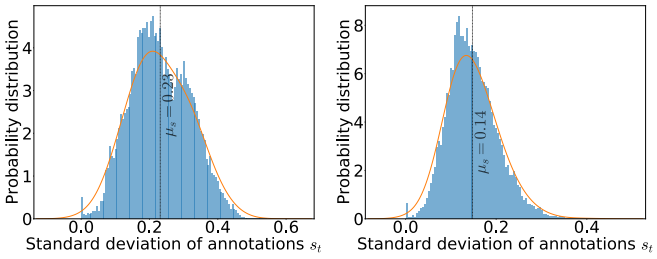
[1]Code for the models and the loss functions introduced are available at https://github.com/sp-uhh/label-uncertainty-ser

To validate the derivation and to further understand the advantages of the t-distribution $\mathcal{L}_{KL}$ (13) over the Gaussian $\mathcal{L}_{KL}$ (14), we plot the $\mathcal{L}_{KL}$ values as a function of varying σ of $\mathcal{Y}_t$, for (13) and (14). We perform this analysis under four different scenarios, by varying parameters $\widehat{\sigma}$ and ν, i) Figure 2a for scenario $\widehat{\sigma} = 0.5$ and $\nu = 6$, ii) Figure 2b for scenario $\widehat{\sigma} = 1.0$ and $\nu = 6$, iii) Figure 2c for scenario $\widehat{\sigma} = 1.0$ and $\nu = 12$, and, iv) Figure 2d for scenario $\widehat{\sigma} = 1.0$ and $\nu = 30$.

From Figure 2, firstly, we see that $\mathcal{L}_{KL}$ behaves differently when the ground-truth $\mathcal{Y}_t$ is modeled as a t-distribution (13), in comparison to the Gaussian assumption (14). Specifically, from Figure 2a, for $\widehat{\sigma} = 0.5$ and $\nu = 6$, we see that the minimum $\mathcal{L}_{KL}$ (13) is achieved only at $\sigma = 0.61$, in contrast to the Gaussian (14) $\widehat{\sigma} = \sigma = 0.5$. While the Gaussian attempts exactly fitting the model to the ground-truth $\sigma = 0.5$, the t-distribution tries to fit on a more relaxed $\sigma = 0.61$ by also considering the reduced degree of freedom $\nu = 6$. This behaviour is similar to that observed during the confidence intervals calculation using a Gaussian and t-distribution [43], where a t-distribution shows relaxation on σ with respect to ν. Moreover, Bishop and Nasrabadi [16] associate this relaxed σ towards the increased robustness of the t-distribution to outliers and sparse distributions.

Secondly, we note that the observed relaxation on σ is dependent on two factors, 1) the standard-deviation of the stochastic outputs $\widehat{\sigma}$, and 2) the degree of freedom of the ground-truth ν. From figures 2a and 2b, we see that, while ν is constant, the relaxation on σ *increases* along with an increase in $\widehat{\sigma}$. At $\widehat{\sigma} = 0.5$ a relaxation of 0.11 is made by t-distribution (13) from 0.5 to 0.61, while a larger relaxation of 0.22 is made for $\widehat{\sigma} = 1.0$. Similarly, from figures 2c and 2d, we see that, while $\widehat{\sigma}$ is constant, as ν increases the relaxation on σ *decreases*. That is, the t-distribution (13) starts behaving similar to that of the Gaussian, inline with literature that states that as the degree of freedom ν of t-distribution increases, the distribution converges into a Gaussian [15], [36], [37]. This is also inline with our initial motivation behind using the t-distribution, which we expected to account for the number of annotators ν while fitting on annotation distribution $\mathcal{Y}$.

From a machine learning and SER perspective, from Figure



(a) For arousal. (b) For valence.

Fig. 3: Distribution of standard deviations s_t in dataset [30]

2, we note several benefits that t-distribution loss term $\mathcal{L}_{KL}$ (13) brings forth in-terms of label uncertainty modeling. Firstly, training on a t-distribution based $\mathcal{L}_{KL}$ (13) leads to training on a relaxed s_t, and thereby can lead to better capturing of the whole ground-truth label distribution. Moreover, the resulting loss function is mathematically more solid than a Gaussian assumption as in [13], [14], when less than *thirty* annotations are available. Secondly, we note that the t-distribution $\mathcal{L}_{KL}$ (13) values are always higher for lower values of σ and $\widehat{\sigma}$, in all cases. This, in comparison to the Gaussian $\mathcal{L}_{KL}$ (14), might lead to larger penalization of the model through the $\mathcal{L}_{KL}$ loss, and may thereby promote better and quicker convergence during training. Finally, the t-distribution $\mathcal{L}_{KL}$ (13) also adapts to different datasets with different number of annotators by considering the number of annotations available during training.

D. Training loss

The proposed uncertainty training loss is formulated as,

$$\mathcal{L} = (1 - \mathcal{L}_{\text{CCC}}(m)) + \mathcal{L}_{\text{BBB}} + \mathcal{L}_{\text{KL}}. \tag{15}$$

Intuitively, $\mathcal{L}_{\text{CCC}}(m)$ (2) optimizes for mean predictions m, $\mathcal{L}_{\text{BBB}}$ (9) optimizes for BBB weight distributions, and $\mathcal{L}_{\text{KL}}$ (13) optimizes for the label distribution $\mathcal{Y}_t$ as a t-distribution.

IV. Experimental Setup

A. Dataset

To validate our model, we use the AVEC'16 [30] version of the RECOLA dataset [20]. In this work, we only utilize the audio signals collected at 16 kHz, from the multimodal signals recorded. The dataset consists of continuous arousal and valence annotations by $a = 6$ annotators at 40 ms frame-rate. As illustrated in Figure 3, in the AVEC'16 [30] dataset, arousal and valence annotations are distributed on average with $\mu_m = 0.01$ and $\mu_m = 0.11$, and $\mu_s = 0.23$ and $\mu_s = 0.14$, respectively, where $\mu_s = \frac{1}{T}\sum_{t=1}^{T} s_t$. This reveals the significant level of subjectivity present in the dataset, where s_t distributions are heavy-tailed with usually high s_t and μ_s. The dataset is divided into speaker disjoint partitions for training, development and testing, with nine 300 s recordings each. As the annotations for the test partition are not publicly available, all results are computed on the development partition.

B. Baselines

To evaluate the performance of our proposed approach, we use MTL- and BBB-based uncertainty models [14]. From [10] we use the perception uncertainty (*MTL PU*) and single-task models (*STL*), and from [14] the model uncertainty (*MU*) and *label uncertainty* (*MU+LU*) algorithms. Similar to [14], for a fair comparison, we reimplemented the baselines from [10], thereby also enabling us to compare the models in-terms of their s estimates, which were not presented in [10]. The method proposed in this work, t-distribution based label uncertainty, will be called henceforth as *t-LU*, for convenience.

C. Choice of hyperparameters

The hyperparameters of the *feature extractor* are fixed as suggested in [38], similar to [44], [45]. The hyperparameters of the *uncertainty layer* are adopted from [14], who show state-of-the-art results in label uncertainty modeling. For example, as *prior distribution* $P(w)$ a simple Gaussian prior with unit standard deviation $\mathcal{N}(0, 1)$ was used. Similarly, the *posterior distribution* $P(w|D)$ is initialized using the same heuristics.

In this work, we assume a Gaussian on $\widehat{\mathcal{Y}}_t$, and noted previously that $n \geq 30$ is required for the assumption to hold. In this light, and keeping the time-complexity in mind, we fixed $n = 30$. For training, we use the Adam optimizer with learning rate 10^{-4}. The batch size used was 5, with a sequence length of 300 frames, 40 ms each. All the models were trained for a fixed 100 epochs. The best model is selected and used for testing when best $\mathcal{L}$ (15) is observed on train partition.

D. Validation measures

To validate the proposed method's *mean* and *standard deviation* estimates, we use $\mathcal{L}_{\text{CCC}}(m)$ and $\mathcal{L}_{\text{CCC}}(s)$ metrics, respectively, widely used in literature [10], [38], [45]. However, $\mathcal{L}_{\text{CCC}}(m)$ and $\mathcal{L}_{\text{CCC}}(s)$ validates mean and standard deviation estimates *separately*. To further jointly validate mean and standard deviation estimates, as label distribution $\widehat{\mathcal{Y}}_t$, we use the $\mathcal{L}_{\text{KL}}$ measure. A similar measure is used in [14], but with a Gaussian assumption on $\mathcal{Y}_t$ and hence can be biased. However, for a fair comparison, we validate all the models in comparison using $\mathcal{L}_{\text{KL}}$ based on their respective distribution assumptions on $\mathcal{Y}_t$, as the models are trained in a similar fashion. The proposed *t-LU* method is validated and trained on the t-distribution $\mathcal{L}_{\text{KL}}$ (13), and the baselines from [10] and [14] are validated and trained on Gaussian $\mathcal{L}_{\text{KL}}$ (14). Nevertheless, from the experiments we also noted that the proposed *t-LU* performs better in-terms of both (13) and (14).

V. Results and Discussion

Table I shows the average performance of the baselines and the proposed model, in terms of their mean m, standard deviation s, and distribution $\widehat{\mathcal{Y}}_t$ estimations, $\mathcal{L}_{\text{CCC}}(m)$, $\mathcal{L}_{\text{CCC}}(s)$ and $\mathcal{L}_{KL}$, respectively. Comparisons with respect to $\mathcal{L}_{\text{CCC}}(s)$ and $\mathcal{L}_{KL}$ are not presented for the STL as this algorithm does not contain uncertainty modeling. Statistical significance is estimated using one-tailed t-test on error distributions, asserting significance for p-values ≤ 0.05, similar to [5].

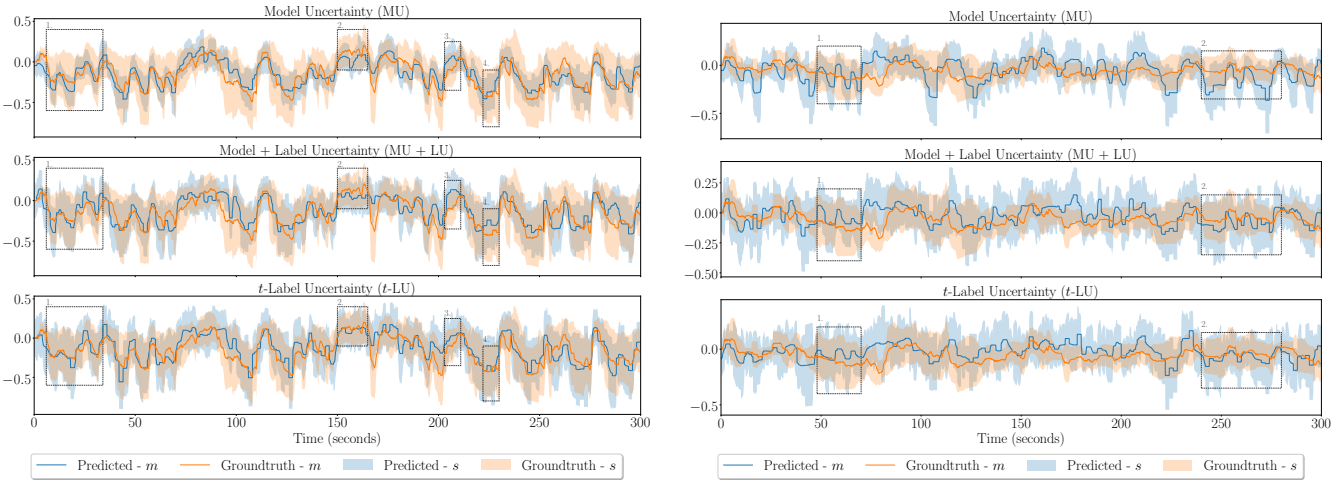
TABLE I: Comparison on mean m, standard deviation s, and label distribution estimations $\mathcal{Y}$, in terms of $\mathcal{L}_{\text{ccc}}(m)$, $\mathcal{L}_{\text{ccc}}(s)$, and $\mathcal{L}_{\text{KL}}$, respectively. Larger CCC indicates improved performance as indicated by $\uparrow$. Lower KL indicates improved performance as indicated by $\downarrow$. ** indicates that the respective approach achieves statistically significant better results than *all* other approaches in comparison. * indicates that it achieves statistically significant better results over *only some* of the approaches in comparison.

(a) For arousal

	$\mathcal{L}_{\text{ccc}}(m)\uparrow$	$\mathcal{L}_{\text{ccc}}(s)\uparrow$	$\mathcal{L}_{\text{KL}}\downarrow$
STL [10]	0.7192	-	-
MTL PU [10]	0.7336	0.2861	0.7965
MU [14]	0.7559	0.0764	0.6900
MU+LU [14]	0.7437	0.3402	0.2576
***t*-LU (proposed)**	**0.7665****	**0.3752****	**0.2349****

(b) For valence

	$\mathcal{L}_{\text{ccc}}(m)\uparrow$	$\mathcal{L}_{\text{ccc}}(s)\uparrow$	$\mathcal{L}_{\text{KL}}\downarrow$
STL [10]	0.3878	-	-
MTL PU [10]	**0.4163**	0.0292	0.9981
MU [14]	0.3248	0.0359	0.6334
MU+LU [14]	0.2831	0.0422	0.4054
***t*-LU (proposed)**	0.3768*	**0.0481***	**0.3914***



(a) For arousal.

(b) For valence.

Fig. 4: Label distribution $\mathcal{Y}_t$ estimation results for a test subject.

A. *Comparison on mean estimates*

In terms of mean estimates $\mathcal{L}_{\text{CCC}}(m)$ of *arousal*, Table Ia shows that the proposed *t*-LU model performs best in comparison with the baselines, with statistical significance. While the *t*-LU model achieves an $\mathcal{L}_{\text{CCC}}(m)$ of 0.7665, the BBB-based baselines, MU+LU and MU, achieve 0.7437 and 0.7559, respectively. This also reveals the superiority of the proposed t-distribution $\mathcal{L}_{KL}$ (13) over the Gaussian $\mathcal{L}_{KL}$ (14), with *t*-LU outperforming MU+LU. Moreover, in [14], it was noted that training on KL loss *with the Gaussian assumption* (14) makes a compromise on $\mathcal{L}_{\text{CCC}}(m)$ performances with improving $\mathcal{L}_{\text{CCC}}(s)$. However, the proposed *t*-LU is free from this compromise with *t*-LU outperforming MU. Finally, we also note that the proposed *t*-LU performs significantly better than the MTL-based uncertainty baseline MTL PU, and also the STL which does not model uncertainty.

In terms of mean estimates $\mathcal{L}_{\text{CCC}}(m)$ of *valence*, concerning Table Ib, it is noted that the the proposed *t*-LU performs significantly better than the BBB-based models, MU+LU and MU. While the *t*-LU model achieves an $\mathcal{L}_{\text{CCC}}(m)$ of 0.3768, the BBB-based baselines, MU+LU and MU, achieve 0.2831 and 0.3248, respectively. Similar to arousal, for valence, we see that *t*-LU is free from compromises on $\mathcal{L}_{\text{CCC}}(m)$, as in [14]. However, the proposed *t*-LU does not improve over the MTL-based baselines, MTL PU and STL. It is a common trend in SER literature that the audio modality inadequately explains the valence dimension of emotion [44]. However, a probable explanation for this is that the MTL-based architectures are generally better than the BBB-based, in-terms of $\mathcal{L}_{\text{CCC}}(m)$ of valence. Results present by Han et al. [10] also show similar behaviour where MTL-based architectures show significant improvements *specifically in mean estimates of valence* [10].

B. *Comparison on uncertainty estimates*

While the proposed *t*-LU achieves significantly improved results over the baselines for mean estimates, especially for the arousal dimension, the main goal of this paper is to aptly capture label uncertainty in emotions, in terms of $\mathcal{L}_{\text{CCC}}(s)$ and $\mathcal{L}_{KL}$. Concerning the uncertainty estimates in arousal, Table Ia shows that the proposed *t*-LU achieves state-of-the-art results, with significant improvements over the baselines. For instance, the MU+LU model, the best performing baseline, achieves a $\mathcal{L}_{\text{CCC}}(s)$ and $\mathcal{L}_{KL}$ of 0.3402 and 0.2576, respectively, while *t*-LU significantly improves by achieving 0.3752 and 0.2349, respectively. This performance, in terms of both the measures, explains the advantage of the t-distribution based KL loss term (13) in label uncertainty modeling. The t-distribution $\mathcal{L}_{KL}$, as seen in Figure 2, promotes the model to fit on a more relaxed

s_t and penalizes more for tighter standard deviations, thereby leading to better capturing the label distribution.

For valence, unlike the $\mathcal{L}_{\text{CCC}}(m)$ performance, Table Ib shows that the proposed t-LU achieves improved performance in-terms of the *uncertainty estimates*, over all the baselines in comparison. It is further noted that t-LU improves with statistical significance over the MTL-based baselines, however improves without statistical significance in comparison with the BBB-based baselines. While the MU+LU, best performing baseline, achieves a $\mathcal{L}_{\text{CCC}}(s)$ and $\mathcal{L}_{KL}$ of 0.0422 and 0.405, respectively, t-LU improves by achieving 0.0481 and 0.3914, respectively. However, the lack of statistical significance over some baselines, as well as the generally rather low performance of all approaches, could be owed to the common observation that speech inadequately explains valence [44].

C. Qualitative analysis on distribution estimation

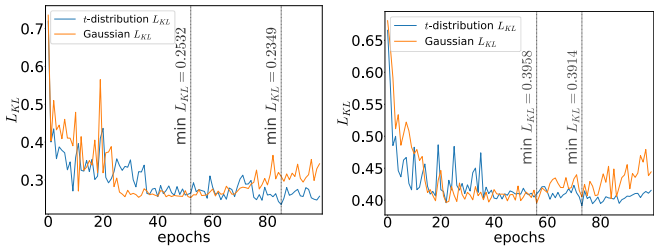
To further validate the results, we plot the mean m and standard deviation s of the estimated distributions for a test subject, seen for arousal in Figure 4a, and valence in Figure 4b. Moreover, some parts of the plots are boxed and numbered to note performance differences. For arousal, in Figure 4a, further backing the results in Table Ia, the proposed t-LU model better captures m and s of the annotation distribution, in comparison with MU and MU+LU. For example, in the box numbered *2* in Figure 4a, the proposed t-LU captures the whole annotation distribution $\mathcal{Y}_t$, resembles the *Ground-truth - s* best. This further highlights the robustness of training on a relaxed σ_t through a t-distribution. Moreover, in box *2*, we also note that t-LU, by best capturing s, also improves notably in terms of the mean estimates m. This can be noted in all boxes *1-4*, where t-LU best captures the whole arousal annotation distribution, by improving on both s and m estimations.

For valence, in Figure 4b, backing results in Table Ib, we note that the proposed t-LU improves significantly in-terms of the mean estimates m, with only small improvements on standard deviation estimates s. This can be seen for instance in box *1*, where t-LU notably improves in terms of m estimations, over MU and MU+LU, while only small improvements in terms of s estimations can be observed. Overall, we note that the proposed speech-based uncertainty model better captures the annotation distribution of arousal than that of valence.

D. Analysis on training loss curve

To further study the advantages of the proposed t-distribution $\mathcal{L}_{KL}$ (13) during the training phase, we compare the testing loss curve of (13) with the Gaussian $\mathcal{L}_{KL}$ in MU+LU (14). The comparison can be seen for arousal in Figure 5a, and for valence in Figure 5b.

For both the arousal and the valence dimension, Figure 5 illustrates two crucial advantages of the proposed t-distribution $\mathcal{L}_{KL}$ (13) as a loss term in the training phase. Firstly, we see that in the initial epochs, before epoch 20, the proposed loss term converges quicker than the Gaussian $\mathcal{L}_{KL}$ (14). This is the result of the proposed $\mathcal{L}_{KL}$ (13) loss term which penalizes more for lower s_t values, in comparison to the



(a) For arousal. (b) For valence.

Fig. 5: Loss curve comparison between Gaussian $\mathcal{L}_{KL}$ (14) and proposed t-distribution $\mathcal{L}_{KL}$ (13).

Gaussian $\mathcal{L}_{KL}$ (14), as seen in Section III-C, thereby achieving faster convergence. Secondly, it is noted that during the later epochs, after 60 epoch, the Gaussian $\mathcal{L}_{KL}$ (14) shows signs of overfitting with increasing testing loss. However, at the same time, the proposed $\mathcal{L}_{KL}$ (13) converges to the best minima during the later epochs. The proposed $\mathcal{L}_{KL}$ achieves minima $\mathcal{L}_{KL}$ at the epoch 85, with $\mathcal{L}_{KL}$ of 0.2349 for arousal and 0.3914 for valence, while the Gaussian achieves a minima well before the later epochs, at epoch 54, with $\mathcal{L}_{KL}$ of 0.2532 for arousal and 0.3958 for valence. The proposed $\mathcal{L}_{KL}$ is free from overfitting in the later stages of training and also learns the optima at this stage. This behaviour can be attributed to the nature of the proposed $\mathcal{L}_{KL}$ which promotes the model to learn a more relaxed s_t, as seen in Section III-C, thereby introducing more regularization to the model, resulting in preventing overfitting and converging on an improved s_t.

VI. Conclusions

Label uncertainty modeling in emotion recognition is commonly approached by assuming a Gaussian distribution on the ground-truth emotion annotations, however with only limited annotations. In contrast, in this work, we assumed a Student's t-distribution on the ground-truth emotion annotations, which also accounts for the number of annotations available. This is the first time in literature an attempt was made to handle the limited emotion annotations available, from a machine learning perspective. For this, we proposed and derived a KL divergence based loss term that aims to capture emotion annotation distribution *as a t-distribution*. The derived t-distribution loss term is also mathematically more sound than the Gaussian assumption, for limited annotations. Subsequently, we showed that the proposed t-distribution loss term leads to training on a relaxed standard deviation, which is adaptable with respect to the number of annotations available. Moreover, we also validated our approach on a publicly available dataset. Quantitative analysis of the results showed that the proposed t-distribution loss term improves over the Gaussian assumption with state-of-the-art results in mean and standard deviation estimations, in-terms of both the CCC and KL divergence measures. Finally, we also showed, through the analysis of the loss curves, that the proposed loss term leads to faster and improved convergence, and is less prone to overfitting.

REFERENCES

[1] G. A. Van Kleef, "How emotions regulate social life: The emotions as social information (easi) model," *Current directions in psychological science*, vol. 18, no. 3, pp. 184–188, 2009.

[2] B. W. Schuller, "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.

[3] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[4] D. Dukes, K. Abrams, R. Adolphs, M. E. Ahmed, A. Beatty, K. C. Berridge, S. Broomhall, T. Brosch, J. J. Campos, Z. Clay *et al.*, "The rise of affectivism," *Nature Human Behaviour*, pp. 1–5, 2021.

[5] K. Sridhar, W.-C. Lin, and C. Busso, "Generative approach using soft-labels to learn uncertainty in predicting emotional attributes," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

[6] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[7] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* IEEE, 2005, pp. 381–385.

[8] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[9] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proceedings of the 25th ACM Int. Conf. on Multimedia*, 2017, pp. 890–897.

[10] J. Han, Z. Zhang, Z. Ren, and B. Schuller, "Exploring perception uncertainty for emotion recognition in dyadic conversation and music listening," *Cognitive Computation*, pp. 1–10, 2020.

[11] K. Sridhar and C. Busso, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *ICASSP - IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2020, pp. 8384–8388.

[12] M. Kumar, E. Sanchez, G. Tzimiropoulos, T. Giesbrecht, and M. Valstar, "Stochastic Process Regression for Cross-Cultural Speech Emotion Recognition," in *Proc. Interspeech 2021*, 2021, pp. 3390–3394.

[13] N. M. Foteinopoulou, C. Tzelepis, and I. Patras, "Estimating continuous affect with label uncertainty," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

[14] N. Raj Prabhu, G. Carbajal, N. Lehmann-Willenbrock, and T. Gerkmann, "End-to-end label uncertainty modeling for speech-based arousal recognition using bayesian neural networks," in *Proc. Interspeech 2022*, Incheon, Korea, September 2022.

[15] S. Kotz and S. Nadarajah, *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.

[16] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[17] H. Fischer, *A history of the central limit theorem: From classical to modern probability theory*. Springer Science & Business Media, 2010.

[18] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 1823–1827.

[19] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[20] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.

[21] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 1022–1040, 2019.

[22] N. Raj Prabhu, C. Raman, and H. Hung, "Defining and Quantifying Conversation Quality in Spontaneous Interactions," in *Comp. Pub. of 2020 Int. Conf. on Multimodal Interaction*, 2020, pp. 196–205.

[23] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[24] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

[25] R. Zheng, S. Zhang, L. Liu, Y. Luo, and M. Sun, "Uncertainty in bayesian deep label distribution learning," *Applied Soft Computing*, 2021.

[26] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, and O. Ronneberger, "A probabilistic u-net for segmentation of ambiguous images," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[27] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. Rezende, and S. A. Eslami, "Conditional neural processes," in *Int. Conf. on Machine Learning*. PMLR, 2018, pp. 1704–1713.

[28] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1613–1622.

[29] M. Abdelwahab and C. Busso, "Active learning for speech emotion recognition using deep neural network," in *2019 8th Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019.

[30] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 3–10.

[31] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, ""of all things the measure is man" automatic classification of emotions and inter-labeler consistency," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–317.

[32] H. M. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *2016 international joint conference on neural networks (IJCNN)*. IEEE, 2016, pp. 566–570.

[33] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," in *Proc. Interspeech*, 2019.

[34] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, "Probability & statistics for engineers & scientists. 2007," *Pearson Cloth, p237*, vol. 1, pp. 67–71, 2017.

[35] J. K. Kruschke, "Doing bayesian data analysis," *Europe's Journal of Psychology*, vol. 7, pp. 778–779, 2010.

[36] C. Villa and S. G. Walker, "Objective prior for the number of degrees of freedom of at distribution," *Bayesian Analysis*, vol. 9, no. 1, pp. 197–220, 2014.

[37] C. Villa and F. J. Rubio, "Objective priors for the number of degrees of freedom of a multivariate t distribution and the t-copula," *Computational Statistics & Data Analysis*, vol. 124, pp. 197–219, 2018.

[38] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," in *ICASSP - Int. Conf. on Acoustics, Speech and Signal Processing*, 2018, pp. 5089–5093.

[39] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[40] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[41] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, 3rd ed., ser. 5. The address of the publisher: MIT Press, 7 2016, vol. 4, ch. 3, pp. 51–77, http://www.deeplearningbook.org.

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[43] D. Rees, "Essential statistics," *American Statistician*, vol. 55, no. 1, 2001, section 9.5.

[44] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Information Fusion*, vol. 68, pp. 46–53, 2021.

[45] P. Tzirakis, A. Nguyen, S. Zafeiriou, and B. W. Schuller, "Speech emotion recognition using semantic information," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6279–6283.

A.3 Leveraging Semantic Information for Speech Emotion Recognition [P5]

Abstract

In large part due to their implicit semantic modeling, self-supervised learning (SSL) methods have significantly increased the performance of valence recognition in speech emotion recognition (SER) systems. Yet, their large size may often hinder practical implementations. In this work, we take HuBERT as an example of an SSL model and analyze the relevance of each of its layers for SER. We show that shallow layers are more important for arousal recognition while deeper layers are more important for valence. This observation motivates the importance of additional textual information for accurate valence recognition, as the distilled framework lacks the depth of its large-scale SSL teacher. Thus, we propose an audio-textual distilled SSL framework that, while having only ∼20% of the trainable parameters of a large SSL model, achieves on par performance across the three emotion dimensions (arousal, valence, dominance) on the MSP-Podcast v1.10 dataset.

Reference

de Oliveira, D. and Raj Prabhu, N. and Gerkmann, T., "Leveraging Semantic Information for Efficient Self-Supervised Emotion Recognition with Audio-Textual Distilled Models", *Proceedings Interspeech*, 3632-3636. 2023. DOI: 10.21437/Interspeech.2023-1758.

Copyright Notice

Authors' Contributions

Danilo de Oliveira is the lead author of this publication, having implemented all algorithms, trained the neural networks utilized in the study, conducted the experimental validation, and authored the manuscript. Navin Raj Prabhu contributed by providing feedback on all methodologies through discussions and assisted in reviewing the manuscript. Timo Gerkmann offered insights on experimental validation and mathematical derivations, in addition to reviewing the manuscript.

Leveraging Semantic Information for Efficient Self-Supervised Emotion Recognition with Audio-Textual Distilled Models

Danilo de Oliveira, Navin Raj Prabhu, Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany

{danilo.oliveira, navin.raj.prabhu, timo.gerkmann}@uni-hamburg.de

Abstract

In large part due to their implicit semantic modeling, self-supervised learning (SSL) methods have significantly increased the performance of valence recognition in speech emotion recognition (SER) systems. Yet, their large size may often hinder practical implementations. In this work, we take HuBERT as an example of an SSL model and analyze the relevance of each of its layers for SER. We show that shallow layers are more important for arousal recognition while deeper layers are more important for valence. This observation motivates the importance of additional textual information for accurate valence recognition, as the distilled framework lacks the depth of its large-scale SSL teacher. Thus, we propose an audio-textual distilled SSL framework that, while having only $\sim$20% of the trainable parameters of a large SSL model, achieves on par performance across the three emotion dimensions (arousal, valence, dominance) on the MSP-Podcast v1.10 dataset.

Index Terms: speech emotion recognition, self-supervised learning, knowledge distillation, paralinguistics, semantics

1. Introduction

Speech signals carry rich information on an individual's emotional states, expressed through both paralinguistic and semantic cues. Backed by the circumplex model [1], research on speech emotion recognition (SER) is typically performed by studying the emotional expressions across multiple dimensions, namely the activation-deactivation dimension (*arousal*), the pleasure-displeasure dimension (*valence*), and the speaker confidence-diffidence dimension (*dominance*). Prior works in literature reveal that certain speech cues carry more information on a particular emotional dimension. For instance, the arousal dimension is well explained by paralinguistic cues [2], and semantic cues are more informative of the valence dimension [3]. This calls for techniques that learn both paralinguistic and semantic cues simultaneously.

The self-supervised learning (SSL) paradigm is a good candidate to fulfill these requirements. Methods from this domain leverage unlabeled data in a pre-training stage as a way of learning robust representations of an input's underlying structure. The pre-training of SSL models is usually succeeded by a fine-tuning stage, in which the model is fed labeled data and trained in a supervised manner for a target downstream task. Following the success of BERT [4] in the natural language processing (NLP) domain, models like wav2vec 2.0 [5] and HuBERT [6] have provided great performance boosts in speech tasks, ranging from automatic speech recognition (ASR) to speaker identification (SI) [7] and SER [8, 9].

While SER systems typically perform well in terms of arousal and dominance, modeling valence from speech signals is a challenging task [10, 2], resulting in a performance gap between arousal/dominance and valence estimation. Prior works have successfully managed to reduce this gap by fine-tuning wav2vec 2.0 and HuBERT for emotion recognition [9]. The improved valence prediction results come from the fact that these transformer-based SSL models can *implicitly* capture semantic content present in speech, along with the paralinguistics [11]. However, such models still underperform on valence compared with models that *explicitly* include semantic information through BERT-encoded features [12]. These findings suggest that the semantic information is not completely modeled by speech-only SSL techniques.

Despite the reduced performance gap between arousal and valence, an important drawback of SSL models is their size: the smallest versions of BERT, wav2vec 2.0, and HuBERT contain 110, 95, and 90 million parameters, respectively, making their use costly or even prohibitive in some cases. Knowledge distillation [13] methods have found success in addressing this issue, in particular in the NLP domain [14, 15]. The goal of these methods is to compress the knowledge acquired by a large network (the teacher) by transferring it to a smaller one (the student). Applied to HuBERT, this framework results in DistilHuBERT [16], with only 25% of the parameters of its teacher.

An analysis of DistilHuBERT suggests that the distilled network is good at encoding paralinguistic information, given the high importance of its representations in the SI task [16]; this indicates potentially good performance in arousal estimation. Moreover, the inner representations of wav2vec 2.0 have an acoustic-linguistic hierarchy [17], where the information embedded in each layer output mutates with network depth, from an acoustic to a semantic nature. These findings further motivate us to investigate how well the distilled model fares in predicting each of the emotional dimensions, especially valence.

In this paper, we make the following contributions: We show that the shallow layers of SSL models are more important for arousal recognition while the deeper layers are more important for valence recognition. We argue that the students in distilled models like DistilHuBERT lack the required depth to properly model valence. Therefore, we show that adding textual information is particularly helpful for distilled SSL networks like DistilHuBERT, considerably more important than for non-distilled large-scale SSL networks. To the best of our knowledge, we are the first in the literature to use distilled SSL models in arousal, valence and dominance modeling and to analyze the implication of the distillation process (i.e., layer selection and compression) towards bridging the gap between valence and arousal estimation.

2. Proposed methodology

The task of emotion recognition is formulated here as follows: given a single-channel audio input containing a spoken utterance $\mathbf{X}_A \in \mathbb{R}^{1\times S}$, where S is the number of samples, we want to estimate three emotional expression scalar values: arousal (Y_a), valence (Y_v) and dominance (Y_d). Our SER model f_{A} should map the input utterance $\mathbf{X}_A$ to an estimate of the three emotion dimensions, simultaneously:

$$\widehat{\mathbf{Y}} = f_{\mathrm{A}}(\mathbf{X}_A), \tag{1}$$

where $\widehat{\mathbf{Y}} = \mathrm{concat}(\widehat{Y}_a, \widehat{Y}_v, \widehat{Y}_d)$.

In the multi-modal case, we also have text as an additional input. The tokenized text is denoted by $\mathbf{X}_T \in \mathbb{N}^{1\times N}$, where N is the number of tokens in the utterance. The audio-textual model mapping is therefore

$$\widehat{\mathbf{Y}} = f_{\mathrm{A+T}}(\mathbf{X}_A, \mathbf{X}_T). \tag{2}$$

Figure 1 illustrates the models. They contain essentially audio and text encoders (SSL_A and SSL_T, respectively), pooling operations and a feed-forward regression head (FFN). The pooling block consists of average- and max-pooling across the sequence dimension. The FFN layer contains two fully-connected layers and a hyperbolic tangent activation function. Its final layer has three output features to estimate $\widehat{\mathbf{Y}}$.

2.1. Audio-only SSL framework

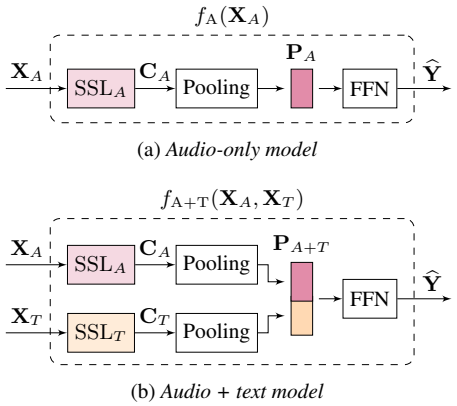
The audio-only system is depicted in Figure 1a. Aiming at extracting features from the time-domain input $\mathbf{X}_A$, we employ pre-trained models as SSL_A, namely wav2vec 2.0, HuBERT and DistilHuBERT. They share the same structure: a convolutional encoder followed by transformer blocks. The convolutional encoder creates frames of latent representations that act as tokens for the creation of contextualized features $\mathbf{C}_A \in \mathbb{R}^{d_A\times K}$ by the transformer encoder, where d_A is the hidden dimension size and K is the number of frames, dependent on the configuration of the convolutional blocks. The transformer encoder is similar to its text counterpart, presented in the next section.

Wav2vec 2.0 uses product quantization to generate a finite set of representations and compares them against the context representations using a contrastive loss. HuBERT, on the other hand, is trained in two separate steps: the first is an offline step that creates discrete pseudo-labels by clustering audio-based features, and the second consists in masked prediction of cluster assignments. Finally, DistilHuBERT is a model obtained by distilling HuBERT through a framework of three prediction heads that aim at estimating HuBERT's 4th, 8th and 12th layers. Though the models make use of largely different pre-training techniques, during fine-tuning/inference their architectures differ essentially in the number of layers and hidden size.

In order to use the extracted representations for our emotion recognition task, the pooling block compresses them to an utterance-level representation $\mathbf{P}_A \in \mathbb{R}^{d_A\times 1}$. This is then fed to the FFN head, which outputs the estimates for arousal, valence and dominance.

2.2. Audio-textual SSL framework

For experiments including text inputs, the token IDs from the tokenized text $\mathbf{X}_T$ pass through an embedding layer, and are then fed to a BERT-like encoder. This class of language models is pre-trained by randomly masking tokens and attempting to predict them using the unmasked ones as context. The architecture is



(a) *Audio-only model*

(b) *Audio + text model*

Figure 1: *System architecture, mapping a spoken utterance* $\mathbf{X}_A$ *(and, in the audio-textual case, its corresponding transcript* $\mathbf{X}_T$*) to a set of arousal, valence and dominance estimates* $\widehat{\mathbf{Y}}$*.*

essentially a sequence of bidirectional transformer layers. The resulting vector $\mathbf{C}_T \in \mathbb{R}^{d_T\times N}$, with d_T being the size of the text hidden dimension, is an embedding enriched by contextual information.

In our experiments, we use TinyBERT [15] as SSL_T, with the intent of minimizing overhead. It is a distilled version of BERT, trained by applying distillation in two steps: pre-training and fine-tuning, aiming at reproducing BERT's capabilities of generalization and downstream task performance, respectively. The authors report 96.8% of the performance of BERT base on an NLP task benchmark, even though the model is 7.5x smaller.

Inspired by [18, 19, 9], we use a simple concatenation method for the fusion of modalities, shown in Figure 1b: the pooled features from each encoder are concatenated in the feature dimension, resulting in $\mathbf{P}_{A+T} \in \mathbb{R}^{(d_A+d_T)\times 1}$. We then pass it through a feed-forward block similar to the audio-only case, only with the number of input features adjusted to incorporate the additional text representations.

3. Implementation details

3.1. Dataset

For training, validation and testing, we use the MSP-Podcast dataset [20], version 1.10. It contains approximately 166 hours of audio extracted from podcast data, labeled at utterance level for arousal, valence and dominance on a scale of 1 to 7. V1.10 features human-labeled transcripts, which serve as our oracle text information. The dataset is split into four partitions: `train`, `development`, `test1` and `test2`. Differently from `test1`, the segments in `test2` originate from podcasts not present in any other partitions. We therefore consider it our *unseen scenarios* set, while `test1` is our *seen scenarios* test data.

3.2. Baseline and proposed models

In our experiments, we propose a framework based exclusively on distilled models and compare it against *base* and *large* SSL_A baselines. As the *base* SSL_A model, we use HuBERT base, which has 12 transformer layers and hidden size $d_A = 768$ [6]. As the *large* SSL_A model, we use the *pruned* w2v2-L-robust from [9], which we denote as w2v2-L-robust(p). It uses the first 12 (out of 24) layers of the original model, with a hidden size $d_A = 1024$. As the proposed distilled SSL_A, we

Table 1: *CCC scores on MSP Podcast (v1.10). Seen scenarios is the test set containing segments from podcasts included in the train or development sets, while the unseen scenarios set contains segments from podcasts not present in any other partition.*

Modality	Model	#Params (M)	Seen Scenarios			Unseen Scenarios		
			Arousal	Valence	Dominance	Arousal	Valence	Dominance
Audio	w2v2-L-robust(p)	165	**0.627**	**0.470**	**0.521**	0.462	**0.263**	0.390
	HuBERT base	95	0.608	0.440	0.486	0.388	0.225	0.344
	DistilHuBERT	**24**	0.622	0.328	0.513	**0.470**	0.169	**0.395**
Audio + Text	w2v2-L-robust(p)	180	**0.619**	**0.560**	0.484	0.469	**0.347**	0.353
	HuBERT base	109	0.613	0.532	0.493	0.461	0.315	**0.396**
	DistilHuBERT	**39**	0.614	0.519	**0.509**	**0.475**	0.333	0.392

use the DistilHuBERT model, that has the same hidden size as HuBERT ($d_A = 768$) but contains only 2 layers. For our audio-textual experiments, as the distilled SSL_T, we employ the 4-layer TinyBERT encoder, with $d_T = 312$.

We make use of the pre-trained SSL models available on Huggingface[1]. When fine-tuning, we follow the usual procedure of freezing the weights of the convolutional encoders present in the speech models. Emotional expression target labels are normalized to $[0, 1]$. We use Adam [21] as the optimizer, with a learning rate of 10^{-5} and early stopping to prevent overfitting. We employ dropout of 0.1 throughout the model. The training data are batched with batch size 16, using buckets sorted by length in order to reduce the amount of padding. The networks are fine-tuned on a single NVIDIA A6000 GPU.

3.3. Loss function

To train the proposed architecture, we use the concordance correlation coefficient (CCC) loss [22], a widely used loss function in SER research [23, 2, 3]. The CCC measures the similarity between two variables and varies between -1 and $+1$, where $+1$ denotes perfect similarity and 0 denotes perfect orthogonality between the variables. For Pearson's correlation coefficient ρ, the CCC between $\mathbf{y}$ and its estimate $\widehat{\mathbf{y}}$ is formulated as

$$\mathcal{L}_{\text{CCC}}(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{2\rho\sigma_{\mathbf{y}}\sigma_{\widehat{\mathbf{y}}}}{\sigma_{\mathbf{y}}^2 + \sigma_{\widehat{\mathbf{y}}}^2 + (\mu_{\mathbf{y}} - \mu_{\widehat{\mathbf{y}}})^2}, \tag{3}$$

where μ, and σ are the mean and standard deviation, respectively, of the corresponding variables. From (3), it can be noted that the CCC takes both the linear correlation and the bias between $\mathbf{y}$ and $\widehat{\mathbf{y}}$ into consideration while quantifying their similarity, hence it is preferred over the Pearson correlation as a loss function for SER. Our training minimizes $1 - \mathcal{L}_{\text{CCC}}(\mathbf{y}, \widehat{\mathbf{y}})$ averaged across the three emotional dimensions.

4. Results and Discussion

4.1. Explicit inclusion of semantic information

We evaluate our audio-only and audio-textual versions of DistilHuBERT along with the HuBERT base and w2v2-L-robust(p) baselines. The results are shown in Table 1. Firstly, from the *audio-only* models, it can be observed that DistilHuBERT manages to match the large SSL baseline w2v2-L-robust(p), for both seen and unseen scenario test data even though it has only $\sim$15% of the total number of parameters. However, the lack of modeling depth and capacity seems to heavily impact valence estimation. The HuBERT base version lags behind in arousal and dominance, but has valence scores closer to those of w2v2-L-robust(p).

[1]https://huggingface.co

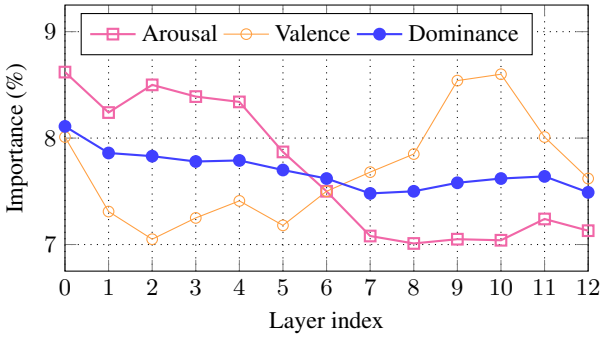


Figure 2: *Normalized importance given to each layer of HuBERT when training in a frozen weights setting. Index 0 corresponds to the output of the convolutional encoder. The subsequent indexes reference each of the transformer encoder's 12 layers.*

Secondly, the *audio-textual* models considerably outperform their audio-only counterparts in terms of valence estimation, even in the case of w2v2-L-robust(p), whose dominance modeling performance decreased, while arousal remained at roughly the same level. The audio-textual versions of the base and distilled HuBERT models either improved or remained at the same level in all cases. Notably, the audio-textual distilled model performs on par with the larger model not only in arousal and dominance but also valence, thus confirming the effectiveness of the explicit inclusion of text, particularly for distilled models.

4.2. Layer importance for the emotional dimensions

As presented in Table 1, the audio-only DistilHuBERT, despite its performance in arousal estimation, performs poorly in terms of valence. To explain this behavior, and also to better understand what kind of information the distilled model learns from the teacher, we proceed by analyzing the relevance of the transformer layers of the pre-trained HuBERT base encoder for arousal, valence and dominance modeling, this time individually. In a similar fashion to [8, 16], we perform a weighted sum of the outputs of the encoder's layers. The weights are normalized to sum to one, and the encoder weights are kept frozen while training the regression head for estimating the target emotional expression. The resulting layer-weighting parameter values found during training are plotted in Figure 2. Index 0 corresponds to the convolutional encoder's outputs, and 1-12 are each of the transformer layers.

Figure 2 reveals that for arousal estimation the first layers up to layer 4 are more important than deeper layers, which hints at the specialization of the first layers on paralinguistic features.

This coincides with the findings of [17] for the similar wav2vec 2.0 model. It is interesting to see that the layer importance follows an opposite trend for valence: Here, lower importance is observed in the first few transformer blocks while deeper layers are more important to model valence. A peak is observed at layers 9 and 10. It raises a concern for the use of DistilHuBERT, since it lacks depth to model the deeper layers. Additionally, its two layers are initialized from the first two of HuBERT. This can explain the distilled model's performance gap between arousal and valence estimations. Furthermore, it also motivates the explicit inclusion of text in SSL models, especially for the distilled models.

4.3. Fine-tuning vs. freezing

Motivated by related works in audio-textual emotion recognition that either fine-tune the SSL encoders' weights [19] or keep them frozen while only training a regression head [18, 9, 12], we run experiments to compare how much each approach helps to improve valence predictions. Figure 3 displays the relative improvement of the valence CCC score over the audio-only case for each of the considered models: large-pruned, base and distilled. "FT" indicates a model with fine-tuned SSL encoders, while "FT⇒FRZ" represents the fusion process used in [9]: a pre-trained speech model is fine-tuned for emotion recognition, then its encoder weights are frozen and used alongside an also frozen text encoder to train a regression head.

We observe in Figure 3 that textual information is particularly helpful for smaller models. This indicates that larger capacity and/or more training data help the model learn some level of semantic information even without explicit textual input. Nevertheless, in all cases the performance is increased by including textual information, confirming the fact that there's still relevant information in the textual input that is not captured by pre-training with audio alone. Furthermore, we observe that fine-tuning brings even larger improvements in all cases; in particular, for the case of the distilled model, valence estimation performance was double that of the audio-only setting in the unseen scenario test set. We therefore hypothesize that fine-tuning the text model helps it focus on modeling the specific semantic information missing from the speech representation.

4.4. Quality of transcripts

When implementing systems that make additional use of text, the audio needs to be transcribed, most practically via an automatic speech recognition (ASR) system. Although practical, ASR models may introduce transcription errors. In order to examine the robustness of the text-informed framework, we run our proposed distilled-only, fine-tuned framework using text transcribed by two ASR systems with different model sizes, and compare it with the model trained on human transcriptions.

The results are presented in Table 2. We can see that the performance is maintained across ASR models, irrespective of their number of parameters. This suggests that the framework is robust with respect to text, as long as the main semantic information is preserved. This claim is further backed by the fact that performance of ASR-generated transcripts is as good as that of human transcripts. It should be noted that the models are run on clean speech; noisy and reverberant conditions are expected to worsen the quality of transcripts and should be addressed by future studies.

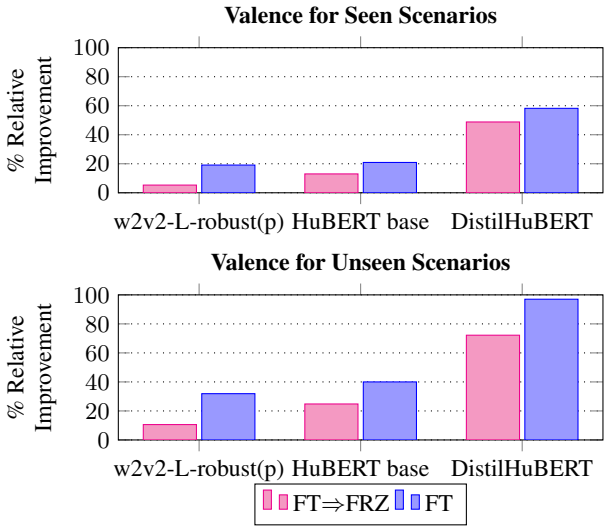


Figure 3: *Relative valence CCC improvement of audio+text models over the audio-only case. "FT⇒FRZ" corresponds to the method used in [9]. "FT" refers to the fine-tuning used in this paper.*

Table 2: *CCC performance of the fine-tuned DistilHuBERT + TinyBERT system with different transcription methods. "A", "V" and "D" denote arousal, valence and dominance, respectively.*

Transcription method	#Params (only ASR)	Seen Scenarios			
		WER	A	V	D
Human	–	–	0.614	0.519	0.509
Whisper base [24]	74M	22.2%	0.620	0.521	0.511
Whisper tiny [24]	39M	24.1%	0.618	0.524	0.510

5. Conclusions

In this study we proposed an audio-textual emotion recognition framework based on distilled models. We highlighted the particular importance of multi-modal audio and text inputs for robust arousal, valence and dominance estimation when using our distilled model. Despite having only ∼20% of the trainable parameters of the largest baseline, the proposed framework's performance is on par with base and large models not only on seen scenarios, but importantly also on unseen scenario data. We investigated the relevance of HuBERT's inner representations to each of the three emotion dimensions and found the initial layers to be more important for arousal modeling, while the deeper layers focus on information instrumental to valence estimation. This analysis further validates the need for text as extra input to distilled networks for improved valence modeling, as these shallow models cannot extract semantic information from speech as easily as their teacher counterparts. Lastly, we confirmed the robustness of our audio-textual network by training it on machine-transcribed audio-text pairs, without loss of performance. Distillation is a promising way to make SSL models more practical, but it is necessary to ensure that both paralinguistic and semantic information is available in order to have robust arousal and valence estimation.

6. References

[1] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[2] N. Raj Prabhu, G. Carbajal, N. Lehmann-Willenbrock, and T. Gerkmann, "End-to-end Label Uncertainty Modeling for Speech-based Arousal Recognition Using Bayesian Neural Networks," in *Interspeech*, Sep. 2022.

[3] P. Tzirakis, A. Nguyen, S. Zafeiriou, and B. W. Schuller, "Speech Emotion Recognition Using Semantic Information," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Jun. 2021.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 1, Jun. 2019, pp. 4171–4186.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, 2020, pp. 12 449–12 460.

[6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE Trans. on Audio, Speech, and Lang. Process. (TASLP)*, vol. 29, pp. 3451–3460, 2021.

[7] N. Vaessen and D. A. Van Leeuwen, "Fine-Tuning Wav2Vec2 for Speaker Recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, May 2022, pp. 7967–7971.

[8] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Interspeech*, Aug. 2021, pp. 3400–3404.

[9] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," Mar. 2022, arXiv:2203.07378 [cs, eess].

[10] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," *IEEE Trans. on Affective Computing*, vol. 13, no. 4, pp. 1959–1972, 2022.

[11] A. Triantafyllopoulos, J. Wagner, H. Wierstorf, M. Schmitt, U. Reichel, F. Eyben, F. Burkhardt, and B. W. Schuller, "Probing Speech Emotion Recognition Transformers for Linguistic Knowledge," in *Interspeech*, Sep. 2022, pp. 146–150.

[12] S. Srinivasan, Z. Huang, and K. Kirchhoff, "Representation Learning Through Cross-Modal Conditional Teacher-Student Training For Speech Emotion Recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, May 2022, pp. 6442–6446.

[13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[14] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," in *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2)- NeurIPS 2019*, Feb. 2020.

[15] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "TinyBERT: Distilling BERT for Natural Language Understanding," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Nov. 2020, pp. 4163–4174.

[16] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "Distilhubert: Speech Representation Learning by Layer-Wise Distillation of Hidden-Unit Bert," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, May 2022, pp. 7087–7091.

[17] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-Wise Analysis of a Self-Supervised Speech Representation Model," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2021, pp. 914–921.

[18] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion Approaches for Emotion Recognition from Speech Using Acoustic and Text-Based Features," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, May 2020, pp. 6484–6488.

[19] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly Fine-Tuning "BERT-Like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition," in *Interspeech*, Oct. 2020, pp. 3755–3759.

[20] R. Lotfian and C. Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings," *IEEE Trans. on Affective Computing*, vol. 10, no. 4, pp. 471–483, Oct. 2019.

[21] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Int. Conf. on Learning Representations (ICLR)*, 2015.

[22] L. I.-K. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, vol. 45, no. 1, p. 255, Mar. 1989.

[23] B. W. Schuller, "Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.

[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," Dec. 2022, arXiv:2212.04356 [cs, eess].

A.4 Ground-truth of affect labels [P6]

Abstract

This chapter presents findings and reflections from an interdisciplinary research project on the automatic detection of affect in group interactions. We begin by examining the inherent subjectivity and ambiguity in how human behavior is perceived during dynamic social interactions, and the resulting uncertainty in behavioral annotations and labels. Drawing on perspectives from both organizational psychology and computer science, we argue that explicitly modeling this ambiguity and subjectivity is essential for advancing the understanding of affect as a social construct. In contrast, subjectivity-agnostic approaches risk oversimplifying complex social phenomena and obscuring meaningful differences in perception. To address this challenge, we explore strategies for modeling label uncertainty and propose a Bayesian Label Distribution Learning (BLDL) framework. Our approach leverages Bayesian neural networks trained on distributions of annotations rather than point estimates, thereby capturing both the diversity of annotator perspectives and the uncertainty that arises from annotation sparsity. Through this subjectivity-aware framework, we demonstrate improved robustness and interpretability in affect recognition compared to traditional methods. We conclude by discussing the broader outcomes and contributions of this research from an interdisciplinary perspective, and by identifying future opportunities where subjectivity-aware modeling can advance not only the study of group affect but also other complex group processes shaped by human perception and interpretation.

Reference

N. Raj Prabhu and V. Begemann and T. Gerkmann and N. Lehmann-Willenbrock, "Ground-Truth of Affect Labels: Accounting for Subjectivity, Ambiguity, and Uncertainty", *Computational group and team dynamics: Forging an interdisciplinary science*, Kozlowski S. W. J. and Hung H. and Lehmann-Willenbrock N. and Salah A. A., Eds. United Kingdom: Oxford University Press, 2025, In Press.

Copyright Notice

Authors' Contributions

Navin Raj Prabhu led the development of the chapter, including the problem formulation, methodological contributions, and the integration of technical arguments into an interdisciplinary narrative. Vanessa Begemann contributed the organizational psychology perspective on affect and played a key role in writing, particularly in making the chapter accessible to both computer scientists and social scientists. Timo Gerkmann contributed through writing support and critical review. Nale Lehmann-Willenbrock provided proofreading, review, and additional interdisciplinary insights.

Ground-Truth of Affect Labels: Accounting for Subjectivity, Ambiguity, and Uncertainty

Navin Raj Prabhu, Vanessa Begemann, Timo Gerkmann, Nale Lehmann-Willenbrock
University of Hamburg, Germany

Abstract

This chapter presents findings and reflections from an interdisciplinary research project on the automatic detection of affect in group interactions. We begin by examining the inherent subjectivity and ambiguity in how human behavior is perceived during dynamic social interactions, and the resulting uncertainty in behavioral annotations and labels. Drawing on perspectives from both organizational psychology and computer science, we argue that explicitly modeling this ambiguity and subjectivity is essential for advancing the understanding of affect as a social construct. In contrast, subjectivity-agnostic approaches risk oversimplifying complex social phenomena and obscuring meaningful differences in perception. To address this challenge, we explore strategies for modeling label uncertainty and propose a Bayesian Label Distribution Learning (BLDL) framework. Our approach leverages Bayesian neural networks trained on distributions of annotations rather than point estimates, thereby capturing both the diversity of annotator perspectives and the uncertainty that arises from annotation sparsity. Through this subjectivity-aware framework, we demonstrate improved robustness and interpretability in affect recognition compared to traditional methods. We conclude by discussing the broader outcomes and contributions of this research from an interdisciplinary perspective, and by identifying future opportunities where subjectivity-aware modeling can advance not only the study of group affect but also other complex group processes shaped by human perception and interpretation.

Keywords: Ground-truth, ambiguity, subjectivity, uncertainty modeling, label distribution learning

Ground-Truth of Affect Labels: Accounting for Subjectivity, Ambiguity, and Uncertainty

Prologue

How Our Collaboration Came About

The subject of group affect and how to account for the inherent subjectivity in affect perceptions and annotations attracted the four co-authors of this chapter for different reasons. Navin, a computer scientist by training, had completed a master's thesis in the area of social signal processing under Hayley Hung and Chirag Raman, working on conversation quality in spontaneous interactions. Along with his interest in social signal processing, Navin also had a keen interest in working on interdisciplinary research questions related to group processes. Timing was of the essence for the collaborative work presented in this chapter. When Navin graduated from his master's degree, he presented his thesis work (Raj Prabhu, Raman, & Hung, 2020) at the 2020 International Conference on Multimodal Interaction (ICMI) workshop (Hung et al., 2020) which was organized by a few contributors to this volume, including Nale. At the end of his talk, Navin casually mentioned that he was on the academic job market. At the time, Nale and Timo were seeking a PhD student with precisely his background to join their recently acquired collaborative project on detecting group affect from audio signals in group conversations. Their search for the ideal candidate had been challenging up to this point.

Nale is an organizational psychologist by training, but frequently integrates methodologies from other disciplines to examine behavioral dynamics in teams and other interaction constellations at work (Lehmann-Willenbrock, 2024). She had previously worked on behavioral expressions of group affect, and continues to be interested in affective convergence and divergence processes in groups and teams (though looking back, the level of granularity in her earlier work on behavioral manifestations of group affect is much less fine-grained than it is now). Driven by curiosity and the quest for more fine-grained analytical methods to study group processes, Nale had been working with

computer scientists in the area of social signal processing for a few years already.

Timo leads the Signal Processing research group at the University of Hamburg, with primary interests in statistical signal processing and machine learning for speech and audio applications. His research has spanned applications such as communication devices, hearing instruments, audiovisual media, and human–machine interfaces. When Nale moved from Amsterdam to Hamburg in search of collaborators in social signal processing, she found Timo—who was keen to expand his work into analyzing group interactions within this field.

Coincidentally, Vanessa also attended the decisive 2020 ICMI workshop, where she presented some of her research on gossip behavior during team meetings. Vanessa completed her PhD on the temporal dynamics and contexts of informal communication dynamics in groups under Nale's supervision. In her research more broadly, she often incorporates human annotations to uncover complex behavioral patterns of groups and understand how individuals interact over time.

Overall, the 2020 ICMI workshop proved to be a turning point for all four of us and inspired us to join forces in different interdisciplinary research projects to understand group dynamics more comprehensively (e.g., Begemann, Hemshorn de Sanchez, Raj Prabhu, Gerkmann, and Lehmann-Willenbrock (2024)). This chapter stems from one such interdisciplinary collaboration.

How our Research Aim Came About, or: "Is there a 'true' label of affect?"

During the early stages of Navin's PhD research, while examining the literature on affect in groups within both computer science and organizational psychology, he raised a significant challenge in one of our project group meetings: the difficulty of defining a "ground truth" for modeling affect, This challenge stems from the inherently subjective and ambiguous nature of affect. *Subjectivity* arises from the inherently personal and interpretive nature of affect. Since affect is often assessed through observable emotional

expressions and quantified via annotations by external observers, the annotated labels inevitably reflect the annotator's subjective interpretations (Alisamir & Ringeval, 2021; Dudzik, Hrkalovic, Hao, Raman, & Tsfasman, 2024; Sridhar & Busso, 2020). *Ambiguity*, on the other hand, emerges from the fact that affective stimuli can evoke multiple, simultaneously present, yet mutually exclusive affective categories (Dudzik et al., 2024; Sethu et al., 2019). Together, subjectivity and ambiguity in labeling affective expressions contribute to a broader, inherent *Uncertainty* in affect annotations. This uncertainty captures both the lack of confidence and the imprecision involved in assigning affective labels to stimuli (Dudzik et al., 2024). Throughout this chapter, we use the term *uncertainty* as an umbrella term that encompasses both the *subjectivity* and *ambiguity* inherent in affect annotations.

When reviewing the literature, it became clear that the essential challenge of accounting for the inherent uncertainty in affect labeling has been frequently overlooked or oversimplified. Common approaches typically involve averaging annotations, applying majority voting, or assuming consensus among annotators, without critically evaluating disagreement or ambiguity in their labels. More importantly, this realization sparked interdisciplinary interest within our team. As computer scientists, Timo and Navin began to explore how uncertainty modeling from computer science could be leveraged to address the issue of subjective and ambiguous affect labels. Meanwhile, Nale and Vanessa observed that organizational psychology research had similarly under-addressed the issue of uncertainty. Responses in organizational psychology often focused on extensively training annotators to improve inter-annotator agreement or simply relied on averaged affect labels for analysis. By comparing our disciplinary perspectives, we recognized a shared gap: the need to more critically engage with the inherent subjectivity and ambiguity in affect labels. This became the foundation for our interdisciplinary collaboration. We saw an opportunity to contribute meaningfully to both fields: providing a deeper understanding of affect in groups for organizational

psychology, and by enhancing affect modeling in computer science through explicitly incorporating label uncertainty. Therefore, our overarching aim was to explore how best to account for the inherent uncertainty in affect labels when capturing and modeling affect.

Affect: A dynamic, multidimensional, and multilevel construct

As an umbrella term for both feeling states and feeling traits (Barsade & Gibson, 2007), affect is a pervasive phenomenon. Whereas feeling traits reflect the individual predisposition to experience more positive or negative affect over time (i.e., positive or negative affectivity), feeling states are dynamic and subject to short-term fluctuations. Feeling states encompass both diffuse mood and discrete emotions that vary along two well-established dimensions: valence (i.e., positive or negative) and arousal, (i.e., level of activation; (Barsade & Gibson, 2007; Lazarus, 1991; J. A. Russell, 1980)). Moods are generally lower in intensity, longer-lasting, and not directly tied to a specific stimulus (Beedie, Terry, & Lane, 2005). Emotions, on the other hand, are more intense, short-lived, and arise in response to specific events or situations (Ekman, 1992). In this chapter, we focus on emotional states and use the terms emotion and affect interchangeably.

Research on affect as emotional states has highlighted three core characteristics: it is a (1) *dynamic*, (2) *multidimensional*, and (3) *multilevel* phenomenon (Barsade & Gibson, 2007; Barsade & Knight, 2015; Lei & Lehmann-Willenbrock, 2015). First, emotional states are dynamic, meaning they are subject to short-term fluctuations in response to events and situational cues. Second, they are multidimensional, encompassing a broad range of both positive (e.g., happiness, joy, excitement) and negative (e.g., sadness, anger, fear, disgust) emotions. These emotions can be mapped along the dimensions of valence and arousal, as defined by the circumplex model (J. A. Russell, 1980) to which we will come to later in this section). Third, emotional states are multilevel phenomena: They can emerge and be observed at the not only at the individual level (e.g., a group member's individual emotion) but also at the group level (e.g., a shared emotional state within a group), particularly during group interactions.

Group interactions provide a critical context not only for the emergence of group-level affect, but also for shaping individual affective experiences. A key

mechanism through which group interactions influences individual affect is emotional contagion (Hatfield, Cacioppo, & Rapson, 1994). Emotional contagion refers to the predominantly automatic and subconscious transmission of emotions between individuals, typically through behavioral mimicry of affective expressions. These affective expressions can be verbal (e.g., humorous remarks or complains), paraverbal (e.g., pitch or volume of voice), or nonverbal (e.g., facial expression or posture). As a form of entrainment, behavioral mimicry functions to align or synchronize individual affective expressions with those of others (Chartrand & Lakin, 2013). Through this bottom-up process, the emotional states of individuals within a group can gradually converge, giving rise to a shared collective group affect - often within a short time frame (Barsade, 2002; Barsade & Knight, 2015; Hatfield et al., 1994). Such bottom-up mechanisms are central to understanding how affect emerges and functions across multiple levels (see for example Barsade and Knight (2015) for an overview) and underscore the critical role of group interactions in shaping both individual and group-level affect. In this chapter, we focus specifically on individual affect during group interactions.

Investigating Affect: From Cross-disciplinary to Interdisciplinary Efforts

Quantifying affect in social interactions

Earlier research in psychology primarily focused on developing theoretical frameworks of individual affective states. Notable examples include the circumplex model, which organizes emotions along two dimensions (i.e., valence and arousal (J. A. Russell, 1980)), and the six basic emotions theory, which identifies six fundamental emotion categories: anger, disgust, fear, happiness, sadness, and surprise (Ekman, 1992). More recent research in organizational psychology has built upon these foundational frameworks by adopting a more process-oriented perspective. In particular, an increasing number of studies have aimed to quantify the dynamic affective processes that unfold within groups and shape group perception and functioning (Barsade & Knight,

2015; Lei & Lehmann-Willenbrock, 2015). To investigate affect and its underlying behavioral dynamics unfolding in group interactions, an increasing number of studies in organizational psychology have employed observational methods and relied on human annotations of affective behavioral expressions(see (Jones, Volet, & Pino-Pasternak, 2021), for an overview). This research has typically focused on verbal (e.g., support statements or complaining; (Lehmann-Willenbrock, Meyers, Kauffeld, Neininger, & Henschel, 2011; Schneider et al., 2018)) and non-verbal (e.g., facial expression and posture; (Bartel & Saavedra, 2000)) behavioral indicators of affect. Verbal affective indicators (e.g., socio-emotional statements) are typically derived from established annotation systems that were developed to capture the functional role of verbal statements during group interactions (e.g., act4teams; (Kauffeld & Lehmann-Willenbrock, 2012)). In contrast, non-verbal behavioral indicators are commonly coded using affect annotation systems in which group-level affect is rated at regular time intervals over the course of an interaction (Bartel & Saavedra, 2000).

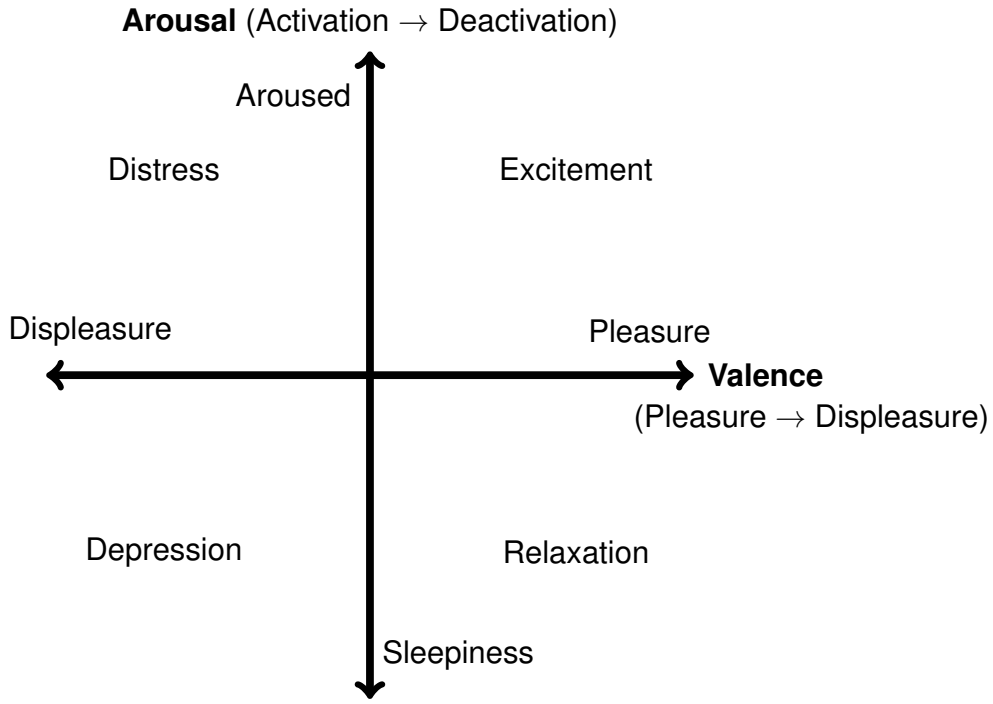


Figure 1. Circumplex Model of Affect illustrating Valence and Arousal dimensions.

To guide human annotations of non-verbal affective expressions, research has traditionally relied on Ekman's six basic emotions theory Ekman and Friesen (1978). This theoretical framework identifies six fundamental emotion categories: anger, disgust, fear, happiness, sadness, and surprise (Ekman, 1992). However, more recent studies have suggested that many affective expressions are inherently ambiguous, lacking clear-cut class boundaries and often fitting multiple categories simultaneously (see J. A. Russell (1980) and Sethu et al. (2019), for further discussions). In light of these limitations, contemporary research on affect has increasingly moved away from discrete emotion categories in favor of continuous representations, most notably through the adoption of the circumplex model (J. A. Russell, 1980). This model maps affective expressions within a two-dimensional space defined by valence (i.e., pleasure-displeasure) and arousal (i.e., activation-deactivation), which are conceptualized as bipolar and orthogonal dimensions. Figure 1 provides a graphical illustration of the circumplex model. The circumplex model provides a more flexible and comprehensive approach to representing affective expression, allowing research to capture affective expressions that may not neatly fit into a single category. It also reflects the gradations and overlaps that characterize natural affective experiences, making it particularly useful for studying affect in complex social settings such as group interactions.

Data-driven affect modeling

In the computer science literature, the field of *Affective Computing* was first introduced by Picard (Picard, 1997) to explore the design of systems and devices capable of recognizing, interpreting, and simulating human affect. Since then, advances in data-driven approaches (S. J. Russell & Norvig, 2015) have significantly accelerated progress in this area. In particular, recent advances in machine learning (Bishop, 2006) and deep neural networks (Goodfellow, 2016) have further enhanced the modeling capabilities of affective systems (for a comprehensive overview, see Calvo, D'Mello,

Gratch, and Kappas (2015); Schuller (2018)). These data-driven models typically employ either an unimodal approach, relying on single behavioral modalities such as either facial expressions (Baltrušaitis, Robinson, & Morency, 2016), speech (Schuller, 2018), or physiological signals only (Alarcao & Fonseca, 2017), or a multimodal approach by integrating multiple modalities (e.g., integrating facial expressions, speech, and movement data; (Tzirakis, Chen, Zafeiriou, & Schuller, 2021; Zeng, Pantic, Roisman, & Huang, 2009). Underpinning these data-driven models theoretical insights from (organizational) psychology - particularly the circumplex model (J. A. Russell, 1980) - have been instrumental in guiding the design and annotation of so called "ground-truth" data.

In the context of data-driven modeling approaches such as machine learning algorithms, establishing reliable ground-truth labels is arguably the most crucial step (Bishop, 2006). "Ground-truth" refers to the accurate and authoritative data or labels that are used as a reference point for training and evaluating the approach. It represents the true outcome or correct label for each instance in a dataset (Bishop, 2006; Goodfellow, 2016). In affective computing, ground-truth labels typically consist of verbal and non-verbal social signals indicative of affect (e.g., socio-emotional statements or prosody of speech). These labels are commonly obtained from human annotators and form the basis for modeling the relationship between social signals and affect. Hence, accurate human annotations are at the core of building automatic affect recognition systems (Schuller, 2018; Sethu et al., 2019).

However, a persistent challenge in the annotation process is the inherent subjectivity and ambiguity involved in perceiving and interpreting affective expressions. This further complicates the concept of a "true" affect label in the objective sense. While Ekman’s influential work (Ekman, 1992; Ekman & Friesen, 1971) proposed that some affective expressions are universal, the perception and interpretation of affective expressions still remains subjective to some extent (Ekman et al., 1987). Research has

demonstrated that the perception and interpretation of affect can be shaped by a variety of factors, including one's own affect (Forgas, 1995) and cultural norms (Porter & Samovar, 1996)). This introduces variability in how different annotators perceive and label the same affective expression.

To address this variability, annotations are often collected from multiple annotators, and consensus on ground-truth is reached using techniques such as averaging scores (Lotfian & Busso, 2019), majority voting (Busso et al., 2008), or computing evaluator-weighted means (EWE) (Grimm & Kroschel, 2005). While these methods aim to increase reliability, they can lead to the loss of valuable information on the inherently subjective nature of affective expressions, potentially masking subtler affective traits (Alisamir & Ringeval, 2021). In real-world applications, capturing the subtle nuances in affective states and achieving robust affect modeling requires more than approximating a consensus-based "true" label. It is equally important to account for the subjectivity inherent in these labels (Alisamir & Ringeval, 2021).

Interdisciplinary efforts

The parallels between data-driven approaches in affective computing and behavioral annotation procedures in organizational psychology have made research on affective processes increasingly interdisciplinary (Gedik, Olenick, Chang, Kozlowski, & Hung, 2023; Lehmann-Willenbrock & Hung, 2024; Maman, Lehmann-Willenbrock, Chetouani, Likforman-Sulem, & Varni, 2024; Nanninga, Zhang, Lehmann-Willenbrock, Szlávik, & Hung, 2017). These interdisciplinary efforts have fueled progress in both fields and have contributed to a more comprehensive understanding of affect in group interactions. Notable examples of such interdisciplinary work include studies by Cohn, Zlochower, Lien, and Kanade (1999); Kanade, Cohn, and Tian (2000); Lucey et al. (2010), which automated the recognition of facial affective states by leveraging foundational psychological research on basic emotions and facial expressions —

particularly the Facial Action Coding System (FACS) (Ekman & Friesen, 1978). These interdisciplinary collaborations have significantly advanced both psychology and computer science research on affect in multiple ways. In psychology, automated affect recognition has reduced the resource-consuming burden of manual annotation and enabled more fine-grained and dynamic analysis of affective expressions (Lewinski, Den Uyl, & Butler, 2014; Onal Ertugrul et al., 2023). For example, whereas the original FACS (Ekman & Friesen, 1978) focused on static, manually coded facial expressions, automated techniques (e.g., Baltrušaitis et al. (2016); Valstar and Pantic (2006)) can now track facial action units continuously over time. This has enabled researchers to investigate the dynamics of affective processes—for instance, tracking the unfolding and attenuation of expressions in clinical contexts (Hamm, Kohler, Gur, & Verma, 2011) as well as predicting moment-to-moment fluctuations in emotional valence and arousal from naturalistic video recordings (Zhang et al., 2024).

In affective computing, automated FACS estimation has provided accurate, labeled datasets on facial action units (AUs)[1] that are critical for large-scale training and evaluation of affect models. Tools like OpenFace and FaceReader have achieved near-human-level accuracy in detecting facial action unites (Baltrušaitis et al., 2016; Den Uyl & Van Kuilenburg, 2015). More recently, generative models have been used to augment AU-labeled datasets to enhance model robustness Sun, Wang, Liu, and Liu (2023). In the context of generative AI, these advancements have enabled the synthesis of realistic facial expressions, improved expression manipulation, and more providede effective dataset augmentation — benefiting applications such as virtual avatars, social robots, and conversational agents (Amini, Lisetti, & Ruiz, 2015; Cohn & De La Torre, 2015).

[1] A facial action unit is a component of facial expression defined in the Facial Action Coding System (FACS), representing the movement of individual or groups of facial muscles (Ekman & Friesen, 1978).

Challenges in collecting labels and ground-truth on affect

While observational methods (in organizational psychology literature) and data-driven techniques (in affective computing literature) have significantly advanced our understanding of the processes of affect, they also present potential challenges, particularly regarding the methods and approaches used to collect labels and ground-truth data for affect. In this chapter, we specifically address the challenge of accounting for the subjectivity and ambiguity inherent in affect labels.

In organizational psychology and related fields, researchers typically aim to minimize subjectivity in behavioral annotations as much as possible. A common strategy for reducing the inherent subjectivity in perceiving and interpreting affective expressions is to establish inter-rater reliability. This involves intensively training at least two independent annotators to apply an annotation system consistently and assessing their agreement on a randomly selected and representative portion of the dataset. Inter-rater reliability is typically quantified using metrics such as Cohen's Kappa or Intraclass Correlation Coefficients (ICC), which indicate the degree of consistency between annotators (Bakeman, McArthur, Quera, & Robinson, 1997). High inter-rater reliability is a widely accepted standard in the field and is often required by publication standards and best-practice recommendations for behavioral annotations in organizational psychology (Güntner, Meinecke, & Lüders, 2023).

However, while high inter-rater reliability ensures that annotators apply the annotation system in a consistent manner, it does not guarantee validity - that is, whether the annotations accurately reflect the expressed affect. For example, two annotators may agree that a particular moment during a meeting reflects "very positive" and "highly activated" group affect according to the applied annotation system. Yet this agreement does not necessarily mean that the annotation system genuinely represents the actual expressed group affect. This disconnect highlights a fundamental limitation of behavioral annotations. Although annotation systems are designed to standardize which observable

behavioral indicators map to which affective states, these mappings are not entirely objective. They may vary across the researchers developing the annotation systems as well as the annotators applying the annotation systems. As a result, even highly reliable annotations can still embed implicit assumptions about the nature of affect, shaped by the theoretical, personal, and cultural lenses of those involved in the annotation process.

In the computer science literature, particularly in the traditional data-driven task of *affect recognition*, researchers address the inherent subjectivity and ambiguity of affect perception by collecting annotations from multiple human annotators, denoted as $\{y_1, y_2, \ldots, y_a\}$ for a annotators (Raj Prabhu et al., 2020; Raman, Raj Prabhu, & Hung, 2023; Ringeval, Sonderegger, Sauer, & Lalanne, 2013). A *ground-truth label*, which serves as the foundation for model training, evaluation, and subsequent analyses, is then derived from these annotations. The most common approach is to compute the simple mean m of the annotators' labels (Abdelwahab & Busso, 2019),

$$m = \frac{1}{a} \sum_{i=1}^{a} y_i. \tag{1}$$

An alternative is the evaluator weighted estimator (EWE) (Grimm & Kroschel, 2005), a weighted mean in which each annotator's contribution is scaled according to their average correlation with the other annotators:

$$m^{\text{EWE}} = \frac{1}{\sum_{i=1}^{a} r^{(i)}} \sum_{i=1}^{a} r^{(i)} \; y_i, \tag{2}$$

where $r^{(i)}$ denotes annotator i's mean correlation with all others. To discount unreliable annotators, any $r^{(i)} < 0$ is set to zero. When all annotators show equal correlation coefficients, the weighting becomes uniform, and the EWE reduces to the unweighted mean $m = m^{\text{EWE}}$.

With these mean- or correlation-based ground truths, the *concordance correlation coefficient* (CCC) (Lawrence & Lin, 1989) has become a widely used loss function during training as well as an evaluation metric in affect recognition (Schuller, 2018). The CCC measures agreement between two variables, such as the ground truth m_t and its

estimate $\widetilde{m}_t$, with values ranging from -1 (perfect disagreement) to $+1$ (perfect agreement). Unlike Pearson's correlation coefficient r, which quantifies only the strength of a linear relationship, the CCC also accounts for mean and scale differences (bias). This makes it particularly well-suited for affect modeling, where both correlation and systematic deviations from the ground truth must be considered. This strategy, in which a single averaged label is treated as the target, is often referred to as the *point-estimate-based approach*, as models are trained and evaluated against a single central value derived from multiple subjective annotations.

However, both m and m^{EWE} capture only the central tendency of a potentially diverse set of subjective annotations, thereby disregarding the subjectivity and ambiguity inherent in individual annotator judgments (Chou & Lee, 2019; Han, Zhang, Ren, & Schuller, 2021; Sridhar & Busso, 2020). For instance, if the annotations ($\{y_1, y_2, \ldots, y_a\}$) follow a normal distribution, the mean m may accurately reflect the group consensus (or central value) due to the symmetry of the data. In contrast, if the annotation distribution is skewed—as often occurs due to high subjectivity and ambiguity in the perception of affective expressions—the mean (m or m^{EWE}) may no longer represent the majority of annotations. Although m^{EWE} places greater weight on more consistent annotators and resembles a form of majority vote, it still overlooks divergent perspectives and less consistent but potentially valid interpretations of affect. As a result, point-estimate-based approaches risk oversimplifying the complexity of human affect perception and under-representing the richness of subjective annotation data.

A further limitation of using mean-based ground-truth is its tendency to suppress less frequent, extreme affective classes within the annotations Sridhar, Lin, and Busso (2021). These outlier annotations may reflect important yet subtle cues that are crucial for understanding complex emotional dynamics in group interactions. By focusing only on the central tendency, both m and m^{EWE}) effectively ignore these expressions, reducing the ecological validity of resulting models.

This narrow treatment of subjectivity introduces at least five limitations. First, it leads to a loss of valuable information from individual annotators, reducing the richness of the dataset and overlooking the variability in affective interpretations. Second, it limits model generalization ability, as models trained on an averaged or singular label (m or m^{EWE}) fail to capture the nuanced nature of affect. This leads to poor performance in diverse real-world scenarios. Third, it risks introducing biases. Assuming a single "ground truth" neglects cultural, contextual, and individual differences in affect perception, often skewing models toward dominant perspectives of certain annotators. Fourth, evaluation metrics based on solely on aggregated labels can give a misleading impression of model performance by failing to account for legitimate disagreements among annotators. Finally, from an ethical standpoint, ignoring subjectivity may also marginalize minority perspectives and inadvertently perpetuate biases in affective computing applications.

In light of these challenges, reliable affect recognition systems - particularly in real-world applications - must move beyond consensus-based modeling of ground-truth labels to explicitly accounting for the inherent subjectivity in those labels (Gunes & Schuller, 2013; Schuller, 2018). Doing so not only improves model robustness but also enables the integration of affect recognition systems in human-in-the-loop solutions and supports the development of algorithms for active learning, co-training, and curriculum learning (Sridhar et al., 2021).

Thus, given the shared challenge of subjectivity and ambiguity in affect labeling, interdisciplinary research between organizational psychology and computer science offers a promising avenue for addressing the resulting label uncertainty. In the following sections, we introduce a data-driven methodology that leverages insights from both fields to more effectively account for this uncertainty and improve the modeling of group affect dynamics.

Research Direction: Accounting for Subjectivity, Ambiguity, and Uncertainty in Affect Labels

Uncertainty modeling: Model uncertainty vs. label uncertainty

In the computer science literature on machine learning and deep learning, subjective and uncertain labels—where a modeling task lacks a well-defined ground-truth, as in the case of affect recognition—are primarily explored under the domain of *Uncertainty Modeling* (Rizos & Schuller, 2019; Tellamekala, Giesbrecht, & Valstar, 2022). Uncertainty modeling in deep learning focuses on quantifying and incorporating uncertainty in model predictions, especially when data or labels are subjective, ambiguous, or incomplete (Kendall & Gal, 2017). Uncertainty modeling in machine learning is typically categorized into two broader categories (Gal & Ghahramani, 2016; Kendall & Gal, 2017). First, *model uncertainty* (or epistemic uncertainty) arises from limited knowledge or insufficient input data and reflects uncertainty in the model's parameters. For example, a model trained on a limited dataset may be uncertain when predicting out-of-distribution samples. Second, *label uncertainty* (or aleatoric uncertainty) arises from inherent noise in the data, such as measurement errors or label ambiguities, and reflects uncertainty in the labels. Affect recognition exemplifies label uncertainty: Different annotators may interpret and label the same expression differently due to its ambiguous nature. Thus, our approach specifically focuses on modeling *label uncertainty* that results from the inherent subjectivity and ambiguity in affect annotations.

Notably, research in affective computing views subjective and ambiguous affect annotations as a cause of label uncertainty, framing it as a cause-and-effect relationship (e.g., Rizos and Schuller (2019); Tellamekala et al. (2022)). Existing research has used multiple terms to represent this scenario of subjective affect annotations, such as 'subjectivity' (Fayek, Lech, & Cavedon, 2016), 'label uncertainty' (Foteinopoulou, Tzelepis, & Patras, 2021), 'soft-labels' (Sridhar et al., 2021) and 'multi-rater labels'

(Chou, Lin, Lee, & Busso, 2022). Henceforth, in this chapter we use the terms 'uncertainty' and 'subjectivity' interchangeably, to represent the subjective and ambiguous nature of affect annotations.

Label uncertainty modeling in affect recognition

Building on the traditional approach of modeling mean-based ground truth (m or m^{EWE}) to represent affect annotations, prior research has sought to address label uncertainty by incorporating an additional ground truth that explicitly represents this uncertainty. One such approach treats label uncertainty as a secondary prediction task, estimating not just the mean-based ground truth but also the annotation distribution (Han et al., 2021; Han, Zhang, Schmitt, Pantic, & Schuller, 2017). Specifically, a multi-task learning (MTL) framework is used to also estimate the unbiased standard deviation s of a annotators as an auxiliary task,

$$s = \sqrt{\frac{1}{a-1}\sum_{i=1}^{a}(y_i - m)^2}. \tag{3}$$

Beyond predicting the unbiased standard deviation s, recent approaches have employed stochastic estimates of affect labels to better capture the uncertainty in affect annotations. Stochastic estimates from deep learning models provide probabilistic outputs that capture uncertainty in predictions (Raj Prabhu, Carbajal, Lehmann-Willenbrock, & Gerkmann, 2022; Rizos & Schuller, 2019; Sridhar & Busso, 2020). These estimates are derived by modeling the distribution of possible outputs, offering insights into the confidence and variability of the model's predictions (Blundell, Cornebise, Kavukcuoglu, & Wierstra, 2015; Gal & Ghahramani, 2016; Kendall & Gal, 2017). For example, Sridhar and Busso (2020) introduced a Monte-Carlo (MC) dropout model to obtain uncertainty estimates from the distribution of stochastic outputs. A similar MC dropout was used by Rizos and Schuller (2019), who proposed a modeling framework that uses uncertainty estimates to identify highly-uncertain samples and perform soft data selection during training.

However, while these stochastic methods model uncertainty, (Rizos & Schuller, 2019; Sridhar & Busso, 2020), they were not explicitly trained on label uncertainty estimates or the distribution of annotations. As a result, they primarily capture *model uncertainty* rather than *label uncertainty*. In other words, these approaches are well-suited to handling uncertainty in input data—such as when encountering unfamiliar or out-of-distribution affective expressions—but they overlook the subjectivity and ambiguity inherent in the perception and human annotation of affect from multiple annotators. As such, we argue that effectively capturing the subjective and ambiguous nature of affective expression requires modeling both model and label uncertainty.

To explicitly model this label uncertainty, a small but growing number of studies in affect modeling, including our own, have proposed training deep learning algorithms using the entire *distribution of affect annotations* as ground-truth rather than relying solely on single-point estimates such as the mean-based ground truth m (Chou et al., 2022; Foteinopoulou et al., 2021; Raj Prabhu, Carbajal, et al., 2022; Raj Prabhu, Lehmann-Willenbrock, & Gerkmann, 2022).

Distribution-based label uncertainty modeling

A central aim of this study is to rethink how affect is understood and modeled by explicitly incorporating the subjectivity and ambiguity present in human annotations. Since affective annotations reflect diverse human perceptions, they naturally form a distribution rather than a single deterministic label (Sridhar et al., 2021). Capturing this distribution is therefore essential for representing the full range of subjective interpretations. In computer science, this idea is formalized as *Label Distribution Learning* (LDL) (Gao, Xing, Xie, Wu, & Geng, 2017), where models are trained directly on label distributions instead of point estimates.

Several implementations of LDL have been explored in affect recognition. For instance, Foteinopoulou et al. (2021) modeled annotations as a univariate Gaussian

within a multi-task learning framework using KL-divergence loss, while Chou et al. (2022) used histogram-based representations of distributions. In our earlier work (Raj Prabhu, Carbajal, et al., 2022), we introduced Bayesian methods (BBB) to approximate Gaussian annotation distributions, thereby bringing uncertainty modeling into LDL for affective data. These prior studies collectively highlight that distribution-based methods provide richer representations of subjective annotations and improve predictive performance compared to point-estimate approaches.

However, existing LDL techniques often rely on simplifying assumptions about the underlying annotation distribution, typically assuming Gaussianity (Chou et al., 2022; Foteinopoulou et al., 2021; Raj Prabhu, Carbajal, et al., 2022). While the Gaussian is mathematically convenient and broadly applicable when the true form of the distribution is unknown, it is highly sensitive to outliers and less reliable when the number of annotations a is small or unevenly distributed (Bishop, 2006; Kotz & Nadarajah, 2004). Given that most affect recognition datasets provide only a limited number of annotations per instance, these limitations are particularly problematic for our overarching goal of faithfully representing subjectivity and ambiguity.

Using a Student's t-distribution for limited and sparse annotations. To overcome these limitations, we extend LDL by modeling annotation distributions with a *Student's t-distribution*, which is more robust to annotation sparsity and outliers (Kotz & Nadarajah, 2004; Walpole, Myers, Myers, & Ye, 2007). Like the Gaussian, the t-distribution is symmetric and bell-shaped, but its heavier tails allow it to better capture extreme values often present in subjective annotations (Bishop, 2006). Crucially, its degrees of freedom parameter ν controls how closely it approximates a Gaussian: as ν increases, the t-distribution converges to the normal distribution (Villa & Rubio, 2018).

In the context of affect recognition, we interpret ν as encoding the number of annotations per instance. This provides a principled way to adapt the distributional model to the availability of annotations, directly addressing one of our study's

contributions: incorporating annotation count into the modeling of uncertainty and subjectivity. As a result, the t-distribution yields more reliable estimates of both central tendency and variability in settings where annotations are sparse (e.g., 3–6 annotators per instance in widely used datasets (Kossaifi et al., 2019; Lotfian & Busso, 2019; Martinez-Lucas, Abdelwahab, & Busso, 2020; Ringeval et al., 2013)). Thus, beyond its statistical advantages, the t-distribution offers a conceptual step toward rethinking affect modeling by linking annotation availability with uncertainty representation.

Deep learning techniques for label distribution learning. While the choice of distribution is critical for modeling subjectivity, an effective framework also requires learning architectures that can represent this distributional uncertainty. Deep learning techniques capable of producing *stochastic outputs* are particularly well-suited for LDL, as they allow the model to approximate predictive distributions rather than single labels. Among such techniques, Bayesian neural networks (BNNs) have been shown to capture label uncertainty more effectively than non-Bayesian probabilistic methods (Gao et al., 2017; Kendall & Gal, 2017).

We therefore adopt *Bayes by Backpropagation* (BBB) (Blundell et al., 2015), a BNN approach that learns distributions over model weights and produces stochastic outputs. These outputs naturally form predictive distributions $\widehat{\mathcal{Y}}$, aligning directly with our overarching aim of learning from the full annotation distribution rather than reducing annotations to point estimates. Compared with alternatives such as ensembles (Liu, Paisley, Kioumourtzoglou, & Coull, 2019), encoder-decoder probabilistic models (Kohl et al., 2018), or neural processes (Garnelo et al., 2018; Tellamekala, Sanchez, Tzimiropoulos, Giesbrecht, & Valstar, 2021), BBB provides both tractability and strong empirical performance for uncertainty-aware LDL tasks. In this way, our deep learning architecture operationalizes the broader conceptual goal of modeling affect as distributions that reflect subjectivity, ambiguity, and annotation sparsity.

Proposed Approach: Bayesian Label Distribution Learning

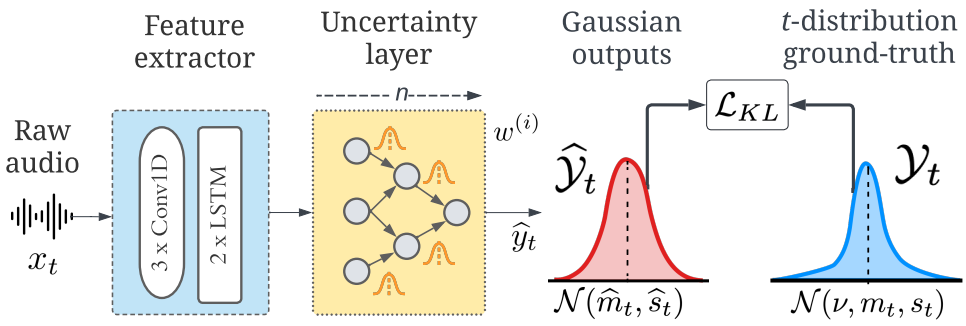


Figure 2. Overview of proposed architecture and loss $\mathcal{L}_{\text{KL}}$. n: number of forward passes. $w^{(i)}$ and $\widehat{y}_t$: stochastically sampled weight and realization of $\widehat{\mathcal{Y}}_t$, at i^{th} forward pass.

Deep learning architecture. To better address the subjectivity and ambiguity inherent in affect annotations, as well as the temporal dynamics of affective expressions, we propose a Bayesian Label Distribution Learning (BLDL) framework. Instead of predicting a single emotion label at each time-frame t from the raw audio input x_t, the BLDL framework estimates the full *emotion annotation distribution* $\mathcal{Y}_t$.

Since the true distributional form of perceived emotions $\mathcal{Y}_t$ is unknown, an initial simplifying assumption is to model it as Gaussian: $\mathcal{Y}_t \sim \mathcal{N}(m_t, {s_t}^2)$, where m_t and s_t represent the mean and variance of the annotations, respectively. This allows the model to capture both the central tendency and the spread of annotations in a principled way. However, in practice, affect annotations are often scarce, unevenly distributed, and may include outliers, making the Gaussian assumption overly restrictive (Bishop, 2006; Kotz & Nadarajah, 2004).

To overcome this limitation, we extend the BLDL framework by modeling $\mathcal{Y}_t$ as a t-distribution with degrees of freedom ν:

$$\mathcal{Y}_t \sim \mathcal{T}(\nu, m_t, {s_t}^2). \tag{4}$$

This formulation offers two advantages. First, the t-distribution is more robust to small sample sizes and outliers, making it a better fit for subjective emotion annotations.

Second, the degrees of freedom ν naturally reflect the number of available annotations, thereby linking data availability to the model's uncertainty representation. In this way, the BLDL framework seeks to approximate the full annotation distribution $\mathcal{Y}_t$, rather than collapsing subjectivity into a single-point estimate.

We design an end-to-end architecture to realize this framework (Figure 2). A feature extractor learns temporal-paralinguistic representations directly from the raw audio waveform x_t. Inspired by Tzirakis, Zhang, and Schuller (2018), it consists of three convolutional layers followed by two stacked long short-term memory (LSTM) layers. On top of this, an *uncertainty layer* estimates the distributional parameters of $\mathcal{Y}_t$. The uncertainty layer is implemented using the Bayes-by-Backprop (BBB) technique (Blundell et al., 2015), consisting of three BBB-based multilayer perceptrons. By producing stochastic outputs $\widehat{y}_t$, this layer approximates the predictive annotation distribution $\widehat{\mathcal{Y}}_t$, thus operationalizing the BLDL framework.

Label distribution loss function. Given the stochastic outputs $\widehat{y}_t$ from the uncertainty layer, label distribution learning is achieved by minimizing a distributional loss that measures the discrepancy between the estimated annotation distribution $\widehat{\mathcal{Y}}_t$ and the ground-truth distribution $\mathcal{Y}_t$. A common choice for this is the Kullback-Leibler (KL) divergence (Gao et al., 2017), denoted as $\mathcal{L}_{\mathsf{KL}}$ in Figure 2. A smaller KL divergence indicates greater similarity between distributions, thereby encouraging the model to produce output distributions that faithfully reflect the subjectively annotated affect.

Under a *Gaussian assumption* for $\mathcal{Y}_t$, the KL divergence reduces to the divergence between two Gaussian distributions. Under the proposed t*-distribution assumption*, we instead compute the KL divergence between $\mathcal{Y}_t \sim \mathcal{T}(\nu, m_t, {s_t}^2)$ and the estimated Gaussian distribution $\widehat{\mathcal{Y}}_t \sim \mathcal{N}(\widehat{m}_t, \widehat{s}_t^{\,2})$. Assuming Gaussianity for $\widehat{\mathcal{Y}}_t$ is justified, since it is constructed from $n \geq 30$ Monte Carlo samples of $\widehat{y}_t$; by the central limit theorem, the resulting empirical distribution converges to Gaussian (Villa & Rubio, 2018; Villa & Walker, 2014).

This setup not only provides a robust match between annotation distributions but also yields a tractable loss. Computing KL divergence directly between two t-distributions involves intractable expectations, whereas divergence between a t-distribution and a Gaussian admits an analytical solution. A detailed derivation of this t-distribution-based KL divergence can be found in our prior work (Raj Prabhu, Lehmann-Willenbrock, & Gerkmann, 2023).

Research Outcomes and Scientific Takeaways

The proposed method advances beyond traditional point estimation by enabling uncertainty estimation in emotion recognition. Rather than providing a single prediction for an emotion label (point estimate) from a given speech input, our approach generates a full emotion distribution, including both the average emotion estimates and its corresponding uncertainty estimates (i.e., deviation). This represents a significant improvement over existing emotion recognition methodologies that rely solely on deterministic, point estimates, such as the E2E Baseline (Tzirakis et al., 2018) or single-task learning technique (STL; Han et al. (2017)). See Table

Early efforts to address this limitation began with multi-task learning (MTL) first introduced by Han et al. (2021). Building upon this, our earlier works introduced Bayesian neural networks (BNN) and label distribution learning (LDL) to further enhance the modeling of affect annotations by explicitly capturing uncertainty. Specifically, we proposed three complementary methodologies for affect recognition: (1) the model uncertainty technique (**MU**), which uses a Bayesian neural network for uncertainty estimation, (2) the label uncertainty technique (G**-LU**), which incorporates label distribution learning under a Gaussian assumption of the emotion annotations, and (3) the t**-LU**, which instead adopts a t-distribution assumption. In this chapter, we systematically evaluate these three approaches to compare their effectiveness and to determine which modeling strategy best captures the inherent subjectivity, ambiguity,

and resulting uncertainty in affect labels for the task of affect recognition and for studying underlying affective processes.

Quantitative analyses of performance

	Arousal		**Valence**	
	$\mathcal{L}_{\text{ccc}}(m) \uparrow$	$\mathcal{L}_{\text{KL}} \downarrow$	$\mathcal{L}_{\text{ccc}}(m) \uparrow$	$\mathcal{L}_{\text{KL}} \downarrow$
E2E Baseline (Tzirakis et al., 2018)	0.770	-	0.361	-
STL (Han et al., 2017)	0.727	-	0.389	-
MTL (Han et al., 2021)	0.740	0.776	**0.420**	0.960
MU (Raj Prabhu, Carbajal, et al., 2022)	0.762	0.675	0.332	0.631
G-LU (Raj Prabhu, Carbajal, et al., 2022)	0.751	0.250	0.301	0.405
t**-LU** (Raj Prabhu et al., 2023)	**0.782***	**0.228***	0.400	**0.386***

Table 1

*Comparison on average m emotion estimation in terms of $\mathcal{L}_{ccc}(m)$, and on uncertainty estimation in terms of $\mathcal{L}_{KL}$ (emotion annotation distribution $\mathcal{Y}_t$ estimation). Larger CCC indicates improved performance as indicated by $\uparrow$. Lower KL indicates improved performance as indicated by $\downarrow$. * indicates that the respective approach achieves statistically significant better results.*

Table 1 presents the performance outcomes of the proposed approach (t-LU) compared to the above discussed techniques. The results demonstrate that the proposed t-LU model excels in both average emotion estimation and uncertainty estimations. More importantly, the results highlight the importance of uncertainty modeling, which not only captures the subjectivity in emotion annotations but also improves modeling of average emotion annotations. Furthermore, our results indicate that incorporating uncertainty modeling in emotion modeling comes with a trade-off between uncertainty estimations and average emotion estimation: *improving* uncertainty estimates ($\mathcal{L}_{\text{KL}}$) results in deteriorates the average emotion estimation ($\mathcal{L}_{\text{CCC}}(m)$). This

becomes evident when comparing the results of MU and G-LU with the performance of Tzirakis et al. (2018). However, our proposed t-LU overcomes this compromise, outperforming the E2E Baseline and other BLDL versions. The improved performance of the proposed t-LU on both average emotion and uncertainty estimation emphasizes the advantage of using the t-distribution based $\mathcal{L}_{KL}$ loss for uncertainty modeling. By promoting the model to fit on a more relaxed s_t, the t-distribution was more robust in capturing the whole label distribution. The fitting on a relaxed s_t leads led increased robustness towards outliers, as noted in Bishop (2006).

Annotation Sparsity and Its Role in Uncertainty Modeling

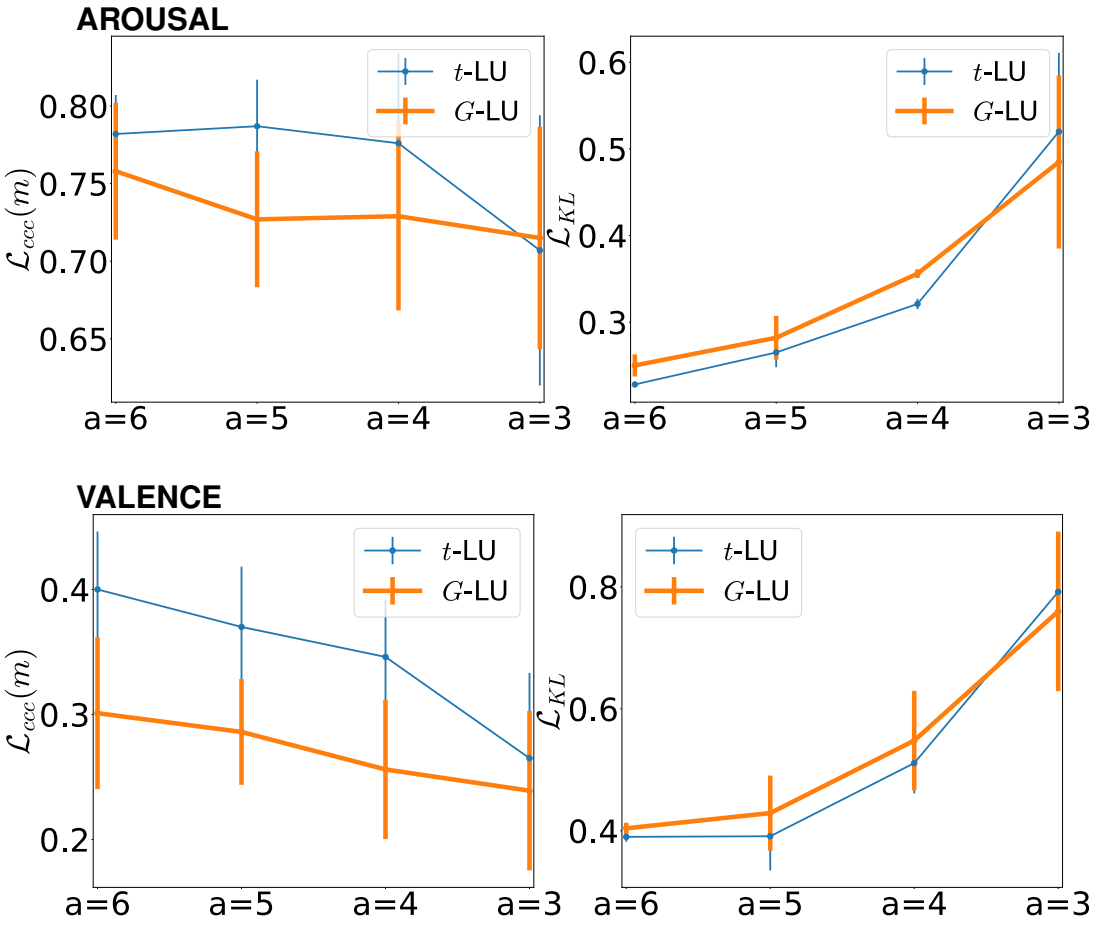


Figure 3. Impact of number of annotations available $a = 6, 5, 4, 3$ on $\mathcal{L}_{\text{ccc}(m)}$ and $\mathcal{L}_{\text{KL}}$.

An important aspect of rethinking how affect is modeled lies in acknowledging not only the variability in annotators' perceptions but also the number of annotations available per instance. The amount of annotation data directly determines how much subjectivity and ambiguity can be represented, yet this factor has received little attention in prior work. Our framework explicitly incorporates this by linking the number of annotations a to the degrees of freedom ν in the t-distribution, thereby adapting the uncertainty estimation to the richness (or sparsity) of the annotation set.

To illustrate this contribution, we conducted experiments where a was systematically varied ($a = 3, 4, 5, 6$). The results, presented in Figure 3, show that when $a \geq 4$, the t-distribution-based t-LU consistently outperforms the Gaussian-based G-LU across datasets. The performance gain was particularly evident at $a = 4$ and $a = 5$, highlighting the robustness of the t-distribution when annotations are limited but not extremely sparse. In contrast, when only $a = 3$ annotations were available, t-LU performed worse than G-LU, reflecting the fact that the t-distribution becomes increasingly uncertain with very few degrees of freedom. These findings demonstrate the broader relevance of our approach: by accounting for how many annotations are available, the t-distribution provides a principled way to represent both subjectivity and annotation sparsity. This moves beyond treating annotations as static ground truth and instead emphasizes the dynamic, data-dependent uncertainty inherent in the social perception of affect.

Discussion

At the core of this work was the acknowledgment that affect annotation is inherently subjective and ambiguous, leading to uncertainty that is often overlooked in conventional modeling approaches. Traditional methods that reduce multiple annotations to a single point estimate risk discarding valuable information about the diversity and disagreement among annotators. By explicitly modeling label distributions and incorporating

uncertainty estimation into the learning process, our aim was to develop a framework that better reflects the complex, multifaceted nature of affect perception. The results of our study confirm this motivation: our Bayesian Label Distribution Learning (BLDL) framework, and in particular the t-LU variant, not only achieves stronger predictive performance but also demonstrates greater robustness under sparse annotations and annotator disagreement. These findings validate the central premise of this research—that embracing rather than ignoring uncertainty is key to advancing the reliability and interpretability of affect recognition systems.

From an interdisciplinary perspective, our contributions are twofold: (1) advancing affect modeling by moving beyond subjectivity-agnostic point estimates toward distribution-based representations of annotator perspectives, and (2) introducing a training methodology that incorporates the number of annotations into the learning process through the loss function. Nevertheless, *several challenges remain*, pointing toward future opportunities for extending subjectivity-aware affect modeling.

Remaining Questions and Future Research Opportunities

First, while our modeling approach enhances the reliability of affect annotations by estimating its uncertainty and acknowledging the inherent subjectivity in affect annotations, the question of validity - whether these annotations actually measure the intended emotions - persists. From a modeling standpoint, this underscores the challenge of construct alignment between data labels and theoretical affective frameworks. Addressing this requires interdisciplinary efforts to align affect labels with established theoretical frameworks of affect, thereby improving construct validity (see also Blanchard et al., in this volume).

Second, our proposed uncertainty modeling approach relied on the audio modality alone. However, our results demonstrate better performance for arousal than for valence. This reflects a well-documented limitation of audio-only emotion recognition, with audio

cues insufficiently explaining the valence dimension of emotions (Sridhar & Busso, 2022; Tzirakis, Chen, et al., 2021; Tzirakis, Nguyen, Zafeiriou, & Schuller, 2021). A likely consequence of this limitation is that, although the t-LU model showed the strongest improvements for valence (Table 1), these gains were not large or consistent enough across samples to reach statistical significance on some evaluation metrics. In other words, the inherent weakness of audio cues for encoding valence dampens the measurable impact of even improved modeling approaches. The performance gap between arousal and valence also highlights a broader modeling challenge: modality-specific expressivity of emotional dimensions. To address this, future work should extend our approach to incorporate multimodal behavioral expressions of affect (e.g., facial, semantic, and physiological cues) to more fully capture the spectrum of affective behavior. This would also improve both predictive performance and interpretability, especially for subtle or ambiguous emotional expressions.

Third, the lack of diverse affect recognition datasets that are accessible to scholars remain a key challenge. This scarcity exacerbates the issue of subjectivity in the ground truths used for training data-driven models. Future interdisciplinary efforts should prioritize the collection of datasets that encompass diverse and naturalistic group interactions across different cultural and organizational contexts, and also include a diverse group of annotators. This would also address the challenge of establishing ecological validity. We are currently taking steps in this direction (Raj Prabhu, Tsfasman, Oertel, Gerkmann, & Lehmann-Willenbrock, 2024).

Finally, it is important to emphasize that while this chapter specifically addresses the uncertainty inherent in affect labels, subjectivity and ambiguity are pervasive across all forms of human social behavior. This includes verbal behavior, such as humor (Romero & Cruthirds, 2006) or gossip (Begemann, Lübstorf, Meinecke, Steinicke, & Lehmann-Willenbrock, 2021), as well as in non-verbal cues such as body gestures and facial expressions in spontaneous interactions (Chang, Lan, Cheng, & Wei, 2020;

Raman, Hung, & Loog, 2022). This reflects a general challenge for computer science and psychology: How can we account for the inherent uncertainty in observed behavioral data? Robust modeling approaches, such as the ones proposed here, are essential not only for affective computing but for building socially intelligent systems that function reliably in real-world contexts. We hope our chapter inspires future interdisciplinary research at the intersection of computer science and psychology to account for the inherent uncertainty in behavioral data and further explore how modeling uncertainty may improve the reliability, validity, and generalizability of annotation schemes and emotion recognition systems.

Interdisciplinary Learning Points

Rethinking the "Understanding of Affect". Taking steps toward the broader interdisciplinary goal of better understanding affect in social interactions, our proposed method for uncertainty modeling offers two main contributions: (1) it refines affect modeling techniques to better represent the inherently subjective and ambiguous nature of annotating affect expressions, and (2) it incorporates the number of annotations available per instance to model uncertainty and subjectivity.

First, we explicitly redefine what it means to “understand affect”. Rather than attempting to recover a single, objective emotional ground truth, we frame understanding as the ability to model the distribution of possible interpretations that different observers might have when perceiving affective expressions. In real-world social interactions, affective expressions are often complex and ambiguous and rarely map cleanly onto discrete, universally agreed-upon emotion classes. Traditional affect modeling techniques typically reduce this complexity by averaging human annotations into a single score or label (e.g., mean ratings), which treats subjectivity and ambiguity as noise. Our proposed modeling technique takes a different approach: Rather than suppressing uncertainty, it seeks to model it directly. By estimating the full distribution of affect

annotations (rather than a single point estimate), our approach captures both the central tendency (e.g., mean) and the variability (e.g., standard deviation) of human perceptions of affective expressions. This offers a richer and more ecologically valid representation of affect.

Second, we emphasize that this is not a model of consensus in the psychological sense of agreement on specific rating points. Instead, our approach models distributional reliability—the statistical stability of the observed distribution as more annotations are collected. Greater numbers of annotations do not necessarily lead to narrow distributions or uniform agreement. Rather, they help us better estimate the shape of the distribution, whether it is tight (reflecting consensus) or wide (reflecting persistent ambiguity). Thus, our method adjusts its confidence not based on the assumption of convergence to a single truth, but based on how well the annotation distribution is supported by the available data. In this way, “understanding affect” in our framework means acknowledging and representing the subjective and context-dependent nature of affective perception, rather than enforcing categorical certainty.

Lessons in Interdisciplinary Collaboration. In addition to the specific contributions to understanding of affect in social interactions, our work also highlights two key learning points for advancing interdisciplinary research at the intersection of organizational psychology and social signal processing more broadly.

First, as with any collaborative endeavor, interdisciplinary research hinges on a shared understanding of the research project - including its constructs, theoretical frameworks, and objectives. In practice, this required extensive and regular discussions to bridge differences in the terminology and the theoretical and practical meaning of concepts across our disciplines. For example, in organizational psychology, labeling observed behavior is commonly refereed to as "coding". When Nale and Vanessa first used this terminology, this initially led to some confusion for Timo and Navin, who as computer scientists have a completely different understanding of what "coding" means

(i.e., programming rather than annotating observed behavioral data). Thus, early phases of our project repeatedly involved translating terminology.

Yet, what initially seemed a straightforward effort to get everyone on the same page became much more fundamental and significantly shaped our project. One of the central concepts in our work, and a focal point of our discussions, was the concept of "ground-truth". While our initial conversations focused on defining the term, they quickly evolved into profound discussions of its philosophical, ethical, and practical meaning. What is "truth" in relation to affect? True to whom and for how long? And on what "grounds"? The supposedly objectivity that the terminology suggests stands in stark contrast to what it practically means in the context of affect recognition: the (average of different) perceptions of behavioral expressions used as a reference for training machine learning models. Our scientific takeaways underscores that there is no single, static ground truth for affect. Instead, we must embrace its variability and model the full range of human interpretations. Our proposed approach reflects this by capturing the entire affect annotation distribution, rather than reducing them to averages.

Second, our work underscores the importance of striking a balance between the contributions from both disciplines. Just as aligning terminology and concepts required extensive discussion, drafting our chapter involved numerous iterations to ensure that both perspectives were adequately represented and the resulting manuscript would be accessible for readers with either type of background. Our feedback loops regularly included comments such as "Are all of these psychological underpinnings really relevant to our specific focus?" (response: no, not all of them) and "This is too 'math-y', can we explain this in simpler terms?" (response: yes, but we still need to retain some equations).

In our experience, the balance of contributions to the different disciplines is also influenced by the maturity of the research area. Our challenge of finding a good balance between the contributions reflected the relatively underdeveloped state of

interdisciplinary research in this domain. In our case, we mutually agreed to draw more heavily from the computer science literature in order to build the necessary foundational work for future interdisciplinary research in this subject area. Nevertheless, we strongly believe that the implications from our work are equally significant for both organizational psychology and computer science literature and point to important avenues for advancing research at the intersection of our fields.

Overall, our experience underscores the importance of open communication, mutual respect, and flexibility in interdisciplinary projects. Establishing a shared understanding, acknowledging the strengths and limitations of each discipline, and finding the right balance of contributions are crucial for successful collaborations. We hope that our lessons learned can serve as a roadmap for future interdisciplinary research on affect dynamics in social interactions.

References

Abdelwahab, M., & Busso, C. (2019, September). Active learning for speech emotion recognition using deep neural network. In *IEEE Int. Conf. on Affective Comp. and Intelligent Interaction.* Cambridge, UK.

Alarcao, S. M., & Fonseca, M. J. (2017). Emotions recognition using eeg signals: A survey. *IEEE Trans. on Affective Computing*, *10*(3), 374–393.

Alisamir, S., & Ringeval, F. (2021). On the evolution of speech representations for affective computing: A brief history and critical overview. *IEEE Signal Proc., Magazine*, *38*, 12–21.

Amini, R., Lisetti, C., & Ruiz, G. (2015). Hapfacs 3.0: Facs-based facial expression generator for 3d speaking virtual characters. *IEEE Trans. on Affective Computing*, *6*(4), 348–360.

Bakeman, R., McArthur, D., Quera, V., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, *2*(4), 357–370. doi: 10.1037/1082-989X.2.4.357

Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In *Winter conf. on applications of computer vision (WACV)* (pp. 1–10).

Barsade, S. G. (2002). The ripple effect: Emotional contagion and its influence on group behavior. *Administrative science quarterly*, *47*(4), 644–675.

Barsade, S. G., & Gibson, D. E. (2007). Why does affect matter in organizations? *Academy of management perspectives*, *21*(1), 36–59.

Barsade, S. G., & Knight, A. P. (2015). Group affect. *Annu. Rev. Organ. Psychol. Organ. Behav.*, *2*(1), 21–46.

Bartel, C. A., & Saavedra, R. (2000). The collective construction of work group moods. *Administrative Science Quarterly*, *45*(2), 197–231. doi: 10.2307/2667070

Beedie, C. J., Terry, P. C., & Lane, A. M. (2005). Distinctions between emotion and

mood. *Cognition and Emotion*, *19*(6), 847–878. doi: 10.1080/02699930541000057

Begemann, V., Hemshorn de Sanchez, C., Raj Prabhu, N., Gerkmann, T., & Lehmann-Willenbrock, N. (2024). Starting on the same note: How pre-discussion small talk and paraverbal synchrony contribute to group entitativity.. (19th Annual INGRoup Conference)

Begemann, V., Lübstorf, S., Meinecke, A. L., Steinicke, F., & Lehmann-Willenbrock, N. (2021, October). Capturing workplace gossip as dynamic conversational events: first insights from care team meetings. *Frontiers in Psychology*, *12*, 725720. doi: 10.3389/fpsyg.2021.725720

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015, July). Weight uncertainty in neural network. In . Lille, France.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., ... Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335–359.

Calvo, R. A., D'Mello, S., Gratch, J. M., & Kappas, A. (2015). *The oxford handbook of affective computing*. Oxford University Press.

Chang, J., Lan, Z., Cheng, C., & Wei, Y. (2020). Data uncertainty learning in face recognition. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* (pp. 5710–5719).

Chartrand, T. L., & Lakin, J. L. (2013). The antecedents and consequences of human behavioral mimicry. *Annual Review of Psychology*, *64*(Volume 64, 2013), 285–308. doi: 10.1146/annurev-psych-113011-143754

Chou, H.-C., & Lee, C.-C. (2019). Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)* (p. 5886-5890). doi:

10.1109/ICASSP.2019.8682170

Chou, H.-C., Lin, W.-C., Lee, C.-C., & Busso, C. (2022, January). Exploiting annotators' typed description of emotion perception to maximize utilization of ratings for speech emotion recognition. In *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP).* Singapore. doi: 10.1109/ICASSP43922.2022.9746990

Cohn, J. F., & De La Torre, F. (2015, 01). 10 automated face analysis for affective computing. In *The oxford handbook of affective computing.* Oxford University Press. doi: 10.1093/oxfordhb/9780199942237.013.020

Cohn, J. F., Zlochower, A. J., Lien, J., & Kanade, T. (1999). Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. *Psychophysiology*, *36*(1), 35–43.

Den Uyl, M., & Van Kuilenburg, H. (2015). Facereader 6: Validation study. *Noldus Information Technology*.

Dudzik, B., Hrkalovic, T. M., Hao, C., Raman, C., & Tsfasman, M. (2024). Indeterminacy in affective computing: Considering meaning and context in data collection practices. In *12th international conference on affective computing and intelligent interaction workshops and demos (aciiw)* (p. 181-185). doi: 10.1109/ACIIW63320.2024.00036

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169–200.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124–129. doi: 10.1037/h0030377

Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., ... Ricci-Bitti, P. E. (1987). Universals and cultural differences in the judgments of

facial expressions of emotion. *Journal of Personality and Social Psychology*, *53*(4), 712–717. doi: 10.1037//0022-3514.53.4.712

Fayek, H. M., Lech, M., & Cavedon, L. (2016, July). Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. In *Ieee int., joint conf., on neural networks (ijcnn).* Vancouver, Canada.

Forgas, J. P. (1995). Mood and judgment: The affect infusion model (aim). *Psychological Bulletin*, *117*(1), 39–66. doi: 10.1037/0033-2909.117.1.39

Foteinopoulou, N. M., Tzelepis, C., & Patras, I. (2021, October). Estimating continuous affect with label uncertainty. In *IEEE Int. Conf. on Affective Comp. and Intelligent Interaction.* Virtual Event.

Gal, Y., & Ghahramani, Z. (2016, June). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *, new york city, ny, usa.*

Gao, B.-B., Xing, C., Xie, C.-W., Wu, J., & Geng, X. (2017). Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, *26*(6), 2825–2838.

Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., ... Eslami, S. A. (2018, July). Conditional neural processes. In *, stockholm, sweden.*

Gedik, E., Olenick, J., Chang, C.-H., Kozlowski, S. W., & Hung, H. (2023). Capturing interaction quality in long duration (simulated) space missions with wearables. *TAC*, *14*(3), 2139-2152. doi: 10.1109/TAFFC.2022.3176967

Goodfellow, I. (2016). *Deep learning.* MIT press.

Grimm, M., & Kroschel, K. (2005, January). Evaluation of natural emotions using self assessment manikins. In *Ieee workshop on automatic speech recognition and understanding* (p. 381-385). doi: 10.1109/ASRU.2005.1566530

Gunes, H., & Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, *31*, 120–136.

Güntner, A. V., Meinecke, A. L., & Lüders, Z. E. (2023). Interaction coding in leadership research: A critical review and best-practice recommendations to measure behavior. *The Leadership Quarterly*, 101751. doi: 10.1016/j.leaqua.2023.101751

Hamm, J., Kohler, C. G., Gur, R. C., & Verma, R. (2011). Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of Neuroscience Methods*, *200*(2), 237-256. Retrieved from https://www.sciencedirect.com/science/article/pii/S016502701100358X doi: https://doi.org/10.1016/j.jneumeth.2011.06.023

Han, J., Zhang, Z., Ren, Z., & Schuller, B. (2021, March). Exploring perception uncertainty for emotion recognition in dyadic conversation and music listening. *Cognitive Computation*, *13*. doi: 10.1007/s12559-019-09694-4

Han, J., Zhang, Z., Schmitt, M., Pantic, M., & Schuller, B. (2017, October). From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proc., of the 25th acm int. conf. on multimedia.* Mountain View, USA.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1994). *Emotional contagion*. Paris, France: Editions de la Maison des Sciences de l'Homme.

Hung, H., Murray, G., Varni, G., Lehmann-Willenbrock, N., Gerpott, F. H., & Oertel, C. (2020). Workshop on interdisciplinary insights into group and team dynamics. In *Proceedings of the international conference on multimodal interaction* (pp. 876–877).

Jones, C., Volet, S., & Pino-Pasternak, D. (2021). Observational research in face-to-face small groupwork: Capturing affect as socio-dynamic interpersonal phenomena. *Small Group Research*, *52*(3), 341–376.

Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Proceedings fourth ieee international conference on automatic face and gesture recognition* (pp. 46–53).

Kauffeld, S., & Lehmann-Willenbrock. (2012). Meetings matter: Effects of team meetings on team and organizational success. *Small Group Research*, *43*(2), 130–158. doi: 10.1177/1046496411429599

Kendall, A., & Gal, Y. (2017, December). What uncertainties do we need in Bayesian deep learning for computer vision? In (Vol. 30).

Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K., ... Ronneberger, O. (2018, December). A probabilistic U-Net for segmentation of ambiguous images. In . Montreal, Canada.

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., ... others (2019). Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Trans., on Pattern Analysis and Machine Intelligence*, *43*(3), 1022–1040.

Kotz, S., & Nadarajah, S. (2004). *Multivariate t-distributions and their applications*. Cambridge University Press.

Lawrence, I., & Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 255–268.

Lazarus, R. S. (1991). *Emotion and adaptation*. New York, NY, US: Oxford University Press.

Lehmann-Willenbrock, N. (2024). Dynamic interpersonal processes at work: Taking social interactions seriously. *Annual Review of Organizational Psychology and Organizational Behavior*, *12*.

Lehmann-Willenbrock, N., & Hung, H. (2024). A multimodal social signal processing approach to team interactions. *Organizational Research Methods*, *27*(3), 477-515. doi: 10.1177/10944281231202741

Lehmann-Willenbrock, N., Meyers, R. A., Kauffeld, S., Neininger, A., & Henschel, A. (2011). Verbal interaction sequences and group mood: Exploring the role of team planning communication. *Small Group Research*, *42*(6), 639–668. doi:

10.1177/1046496411398397

Lei, Z., & Lehmann-Willenbrock, N. (2015). Dynamic affect in team meetings: An interpersonal construct embedded in dynamic interaction processes. In *The cambridge handbook of meeting science* (pp. 456–480).

Lewinski, P., Den Uyl, T. M., & Butler, C. (2014). Automated facial coding: Validation of basic emotions and facs aus in facereader. *Journal of neuroscience, psychology, and economics*, *7*(4), 227.

Liu, J., Paisley, J., Kioumourtzoglou, M.-A., & Coull, B. (2019, December). Accurate uncertainty estimation and decomposition in ensemble learning. In . Vancouver.

Lotfian, R., & Busso, C. (2019, December). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Trans. on Affective Computing*, *10*(4), 471-483. doi: 10.1109/TAFFC.2017.2736999

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer society conference on computer vision and pattern recognition-workshops* (pp. 94–101).

Maman, L., Lehmann-Willenbrock, N., Chetouani, M., Likforman-Sulem, L., & Varni, G. (2024). Modeling the interplay between cohesion dimensions: A challenge for group affective emergent states. *IEEE Trans. on Affective Computing*, *15*(3), 1526-1538. doi: 10.1109/TAFFC.2024.3349910

Martinez-Lucas, L., Abdelwahab, M., & Busso, C. (2020, October). The MSP-conversation corpus. In *Interspeech.* Shanghai, China.

Nanninga, M. C., Zhang, Y., Lehmann-Willenbrock, N., Szlávik, Z., & Hung, H. (2017). Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry. In *Int. Conf. on Multimodal Interaction (ICMI)* (pp. 206–215). New York, New York, USA.

Onal Ertugrul, I., Ahn, Y. A., Bilalpur, M., Messinger, D. S., Speltz, M. L., & Cohn, J. F. (2023). Infant afar: Automated facial action recognition in infants. *Behavior research methods*, *55*(3), 1024–1035.

Picard, R. W. (1997). *Affective computing*. Cambridge, MA, USA: MIT Press.

Porter, R. E., & Samovar, L. A. (1996). Chapter 17 - cultural influences on emotional expression: Implications for intercultural communication. In P. A. Andersen & L. K. Guerrero (Eds.), *Handbook of communication and emotion* (p. 451–472). San Diego: Academic Press. doi: 10.1016/B978-012057770-5/50019-9

Raj Prabhu, N., Carbajal, G., Lehmann-Willenbrock, N., & Gerkmann, T. (2022, September). End-to-end label uncertainty modeling for speech-based arousal recognition using Bayesian neural networks. In *Interspeech.* Incheon, Korea.

Raj Prabhu, N., Lehmann-Willenbrock, N., & Gerkmann, T. (2022, October). Label uncertainty modeling and prediction for speech emotion recognition using t-distributions. In *IEEE Int. Conf. on Affective Comp. and Intelligent Interaction.* Nara, Japan.

Raj Prabhu, N., Lehmann-Willenbrock, N., & Gerkmann, T. (2023). End-to-end label uncertainty modeling in speech emotion recognition using bayesian neural networks and label distribution learning. *IEEE Trans. on Affective Computing*, 1-14. doi: 10.1109/TAFFC.2023.3283595

Raj Prabhu, N., Raman, C., & Hung, H. (2020, September). Defining and Quantifying Conversation Quality in Spontaneous Interactions. In *Companion Publiciton of the Int. Conf. on Multimodal Interaction (ICMI).*

Raj Prabhu, N., Tsfasman, M., Oertel, C., Gerkmann, T., & Lehmann-Willenbrock, N. (2024). Dynamics of collective group affect: Group-level annotations and the multimodal modeling of convergence and divergence. Retrieved from `https://arxiv.org/abs/2409.08578` (Under Review)

Raman, C., Hung, H., & Loog, M. (2022). Social processes: Self-supervised

meta-learning over conversational groups for forecasting nonverbal social cues. In *European conference on computer vision* (pp. 639–659).

Raman, C., Raj Prabhu, N., & Hung, H. (2023, January). Perceived conversation quality in spontaneous interactions. *IEEE Trans. on Affective Computing*, 1-13. doi: 10.1109/TAFFC.2023.3233950

Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013, April). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Ieee int. conf. and workshops on automatic face and gesture recognition (fg).* Shanghai, China.

Rizos, G., & Schuller, B. (2019, May). Modelling sample informativeness for deep affective computing. In *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP).* Brighton, UK. doi: 10.1109/ICASSP.2019.8683729

Romero, E. J., & Cruthirds, K. W. (2006). The use of humor in the workplace. *Academy of Management Perspectives*, *20*(2), 58-69. doi: 10.5465/amp.2006.20591005

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Russell, S. J., & Norvig, P. (2015). *Artificial intelligence: A modern approach* (3rd ed.). Upper Saddle River, NJ: Pearson.

Schneider, K., Klünder, J., Kortum, F., Handke, L., Straube, J., & Kauffeld, S. (2018). Positive affect through interactions in meetings: The role of proactive and supportive statements. *Journal of Systems and Software*, *143*, 59–70.

Schuller, B. W. (2018, April). Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, *61*(5), 90–99.

Sethu, V., Provost, E. M., Epps, J., Busso, C., Cummins, N., & Narayanan, S. (2019). The ambiguous world of emotion representation. *arXiv preprint arXiv:1909.00360*.

Sridhar, K., & Busso, C. (2020, May). Modeling uncertainty in predicting emotional attributes from spontaneous speech. In *IEEE Int. Conf. on Acoustics, Speech and*

Signal Process. (ICASSP). Barcelona, Spain.

Sridhar, K., & Busso, C. (2022, June). Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech. *IEEE Transactions on Affective Computing*, 1-17. doi: 10.1109/TAFFC.2022.3187336

Sridhar, K., Lin, W.-C., & Busso, C. (2021, October). Generative approach using soft-labels to learn uncertainty in predicting emotional attributes. In *IEEE Int. Conf. on Affective Comp. and Intelligent Interaction* (pp. 1–8). Virtual Event.

Sun, Y., Wang, T., Liu, J., & Liu, Z. (2023). Gan-au: Facial action unit detection via data augmentation with generative adversarial networks. In *Proceedings of the ieee/cvf winter conference on applications of computer vision.*

Tellamekala, M. K., Giesbrecht, T., & Valstar, M. (2022, July). Dimensional affect uncertainty modelling for apparent personality recognition. *IEEE Trans. on Affective Computing*. doi: 10.1109/TAFFC.2022.3189974

Tellamekala, M. K., Sanchez, E., Tzimiropoulos, G., Giesbrecht, T., & Valstar, M. (2021, September). Stochastic Process Regression for Cross-Cultural Speech Emotion Recognition. In *Interspeech.* Brno. doi: 10.21437/Interspeech.2021-610

Tzirakis, P., Chen, J., Zafeiriou, S., & Schuller, B. (2021). End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, *68*, 46–53.

Tzirakis, P., Nguyen, A., Zafeiriou, S., & Schuller, B. W. (2021, June). Speech emotion recognition using semantic information. In *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP).* Toronto.

Tzirakis, P., Zhang, J., & Schuller, B. W. (2018, April). End-to-End Speech Emotion Recognition Using Deep Neural Networks. In *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP).* Calgary, Canada.

Valstar, M., & Pantic, M. (2006). Fully automatic facial action unit detection and temporal analysis. In *2006 conference on computer vision and pattern recognition workshop*

(cvprw'06) (pp. 149–149).

Villa, C., & Rubio, F. J. (2018). Objective priors for the number of degrees of freedom of a multivariate t distribution and the t-copula. *Computational Statistics & Data Analysis*, *124*, 197–219.

Villa, C., & Walker, S. G. (2014). Objective prior for the number of degrees of freedom of at distribution. *Bayesian Analysis*, *9*(1), 197–220.

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2007). *Probability & statistics for engineers and scientists*. Pearson Education.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 39-58. doi: 10.1109/TPAMI.2008.52

Zhang, J., Sato, W., Kawamura, N., Shimokawa, K., Tang, B., & Nakamura, Y. (2024). Sensing emotional valence and arousal dynamics through automated facial action unit analysis. *Scientific Reports*, *14*(1), 19563.

B

Related Abstract Presentations

B.1 Small Talk, Synchrony, and Entitativity [P7]

Abstract

In our interdisciplinary study, we investigate the influence of pre-discussion small talk on group entitativity and examine synchrony as a mediating mechanism. Our preliminary findings provide novel insights into the intricate processes in group interactions that contribute to groups perceiving themselves as units.

Reference

V. Begemann and C.S. Hemshorn de Sanchez and N. Raj Prabhu and T. Gerkmann and N. Lehmann-Willenbrock, "Starting on the Same Note: The Effect of Pre-Discussion Small Talk on Group Entitativity through Synchrony", *19th Annual INGRoup Conference*, Charlotte, North Carolina, USA, July, 2024.

Copyright Notice

Authors' Contributions

This work was a collaboration between the Organizational Psychology group and the Signal Processing group, at the University of Hamburg. Vanessa Begemann and Clara S. Hemshorn de Sanchez, as organizational psychology experts, were the lead contributors, responsible for setting up the experiment, collecting the group interaction data, and conducting the analysis and experimental validation. Navin Raj Prabhu, as the signal processing expert, developed the algorithms for extracting synchrony- and mimicry-based features and provided these for subsequent analysis. Timo Gerkmann, also as the signal processing expert, offered insights on algorithm design and assisted in reviewing the manuscript. Nale Lehmann-Willenbrock offered valuable theoretical feedback and participated in the manuscript review.

Starting on the Same Note: The Effect of Pre-Discussion Small Talk on Group Entitativity through Synchrony

Vanessa Begemann, Clara Sofie Hemshorn de Sanchez, Navin Raj Prabhu, Timo Gerkmann, and Nale Lehmann-Willenbrock

Abstract

In our interdisciplinary study, we investigate the influence of pre-discussion small talk on group entitativity and examine synchrony as a mediating mechanism. Our preliminary findings provide novel insights into the intricate processes in group interactions that contribute to groups perceiving themselves as units.

Keywords: entitativity, synchrony, small talk, group dynamics, social signal processing

Purpose

Entitativity, which reflects the perception of a group as a cohesive unit, is a fundamental characteristic of groups (Blanchard et al., 2020; Campbell, 1958). It is the starting point for central group outcomes, such as cohesion, identification, and satisfaction (Blanchard et al., 2020), highlighting its importance for group functioning. It is influenced by similarity and interactivity between individuals (Blanchard et al., 2020). Yet, while the influence of formal interactions, such as meetings, on entitativity has been documented in previous literature (Blanchard et al., 2022; Blanchard & McBride, 2020), the role of informal communication in fostering a sense of unity remains empirically underexplored (Blanchard & Allen, 2023). This is problematic because informal interactions constitute a significant amount of overall communication at work (Koch & Denner, 2022; Whittaker et al., 1994) and have become increasingly important with the rise of remote and hybrid work arrangements (Blanchard, 2021; Bleakley et al., 2021).

One central type of informal interactions is pre-discussion small talk (Allen et al., 2014; Mirivel & Tracy, 2005). Prior to a more formal discussion, the air often fills with seemingly inconsequential chatter about the weather, coffee preferences, or the mundane details of daily life. These moments of small talk, characterized by atopical and minimally informative communication (Holmes, 2003; Methot et al., 2021), are ubiquitous informal interactions at work (Keyton et al., 2013). Although its substance is inconsequential, small talk can have substantial consequences for groups. It helps individuals socialize before meetings, transition to more serious discussions, enhances wellbeing, creates positive group climates, and facilitates social bonds and a sense of belonging (Allen et al., 2014; Holmes, 2003; Holmes & Marra, 2004; Methot et al., 2021; Mirivel & Tracy, 2005; Molinsky, 2013). In fact, Coupland et al. (1992)

argue that establishing relationships is a key function of small talk. Overall, being a form of ritualized interaction – which signals inclusion and unifies individuals (Collins, 2005) – , we therefore argue that pre-discussion small talk plays an important role in shaping group entitativity.

Moreover, the role of small talk as a precursor to enhanced entitativity raises questions about how these informal interactions contribute to the perception of a group as a unified entity. Following the notion that similarity is a key determinant of entitativity (Blanchard et al., 2020), we argue that synchrony is a dynamic mechanism through which small talk enhances entitativity. Related to similarity, synchrony refers to the temporal coordination of actions between two or more individuals and can involve verbal (e.g., spoken words), paraverbal (e.g., pitch of voice), and non-verbal cues (e.g., body movement; Lehmann-Willenbrock & Hung, 2023). Indeed, previous research points to its benefits for social cohesion, for example, by synchronized movement and arousal (Jackson et al., 2018) or language convergence (Van Swol & Kane, 2019). Thus, we argue that synchrony, in particular synchrony in paraverbal behavior, shapes group entitativity. Furthermore, as small talk is a ritualized interaction that specifically aims at "greasing the social wheels" and easing into more complex discussions (Coupland et al., 1992; Holmes, 2003; Holmes & Marra, 2004; Methot et al., 2021), we suggest that this is characterized by a stronger synchronization of group members. Therefore, we propose that synchrony mediates the effect of pre-discussion small talk on group entitativity.

By systematically examining how pre-discussion small talk influences perceptions of entitativity and identifying synchrony as a potential mediating factor, our interdisciplinary study offers two important theoretical contributions. First, our study extends the understanding of entitativity by focusing on group-level entitativity and incorporating the role of informal

interactions, thus broadening the scope of antecedents of entitativity. Second, we provide important insights into the mechanisms through which informal interactions facilitate a sense of entitativity within groups, contributing to both group research and social signal processing literature. Overall, we provide a novel examination of the intricate processes that contribute to the perception of a group as a unified entity.

Method

Participants and Procedure

We recruited $N = 30$ groups of three students each, all attending the same university. Participants (58.9% female, 34.4% male, 6.7% non-binary) were on average 24.30 years old ($SD = 4.09$). The students were invited for group discussion about potential decarbonization measures at their university. Half of the student groups were randomly selected to engage in small talk prior to the discussion. They were asked to quickly (max. three minutes) talk about whether and how they drank their coffee today, and whether they are happy with today's weather. During the group discussion, all groups were asked to discuss different fictional, controversial decarbonization measures, such as the elimination of disposable coffee cups at the university or restrictions on conference attendance based on travel limitations. The twelve different decarbonization measures were presented on discussion cards, allowing groups the autonomy to select another card after completing discussing one measure or skip to a different measure as desired. The group discussions were video-recorded, lasting on average 18 minutes long ($SD = 4$ minutes), with durations ranging from 11 to 24 minutes. Before and after the discussion, the participants were asked to fill out questionnaires. Data collection for the study is still ongoing.

Measures

Entitativity

To measure entitativity, we used the following three items from Blanchard et al. (2020): "We are a unit.", "We are a group", "We feel like a group to me" ($\alpha = .75$). The items were measured on a 7-point response scale (1 = *not at all* to 7 = *completely*), indicating the extent of participants' agreement with each statement. Given our focus on group-level entitativity, we evaluated the appropriateness of aggregating individual responses to the group level by assessing within-group agreement using the $r_{WG(j)}$ statistic. According to our previously statistically determined ideal cutoff value of $r_{WG(j)} = .87$, the observed mean within-group agreement $r_{WG(j)} = .63$ was not significantly greater than chance. However, the observed $r_{WG(j)}$-value ranged from 0 to .98, with a median of .75. Thus, according to the cutoff value of $r_{WG(j)}. = 70$ proposed by Bliese (2000), and following the heuristics suggested by LeBreton and Senter (2008), these findings indicate that at least half of the groups showed strong within-group agreement. Thus, we aggregated responses to reflect group-level entitativity.

Synchrony

We measured synchrony of the group by analyzing the audio recordings from group discussions, with a particular focus on the synchrony of pitch among group members. Pitch, which reflects the tonal height or depth of speech, serves as a quantifiable measure of vocal expression. Following Nanninga et al. (2017), we pre-processed the audio signals by extracting its pitch contours, prior to the synchrony calculation. The pitch contours were calculated using the YIN algorithm which is an autocorrelation based method for fundamental frequency estimation. Next, considering the observed average length of speaking turns in our data ($M = 12$ seconds, $SD = 14$ seconds), we segmented the audio stream of each discussion into 30-second

windows. Subsequently, informed by the SyncPy Package (Varni et al., 2015), we developed a code to compute synchrony-based Pearson's correlation coefficients. These coefficients were calculated for the pitch contours of each dyadic pairing within the group (e.g., member A – B, B – C, and A – C) over each 30-second window, without incorporating any time lag. Finally, we calculated a mean Pearson's coefficient per group by averaging the coefficients across all dyadic pairings within a group and across all windows. This aggregate value represents the overall synchrony for each group discussion, with higher values indicating stronger synchrony among group members.

Preliminary Results and Discussion

To test our hypotheses, we conducted a mediation analysis using R version 4.2.2 (R Core Team, 2022), with small talk as the predictor, synchrony as the mediator, and entitativity as the outcome variable. We observed a significant positive direct effect of small talk on entitativity ($B = 0.53$, $p < .001$), indicating that groups engaging in pre-discussion small talk reported higher levels of group entitativity ($M = 5.70$, $SD = 0.39$) compared to those that did not engage in pre-discussion small talk ($M = 5.24$, $SD = 0.37$). Additionally, we observed a significant positive effect of synchrony on entitativity ($B = 1.78$ $p < .001$). Contrary to our theorizing, the effect of small talk on synchrony was significant but negative ($B = -0.04$, $p = .042$), indicating that groups engaging in pre-discussion small talk exhibited lower synchrony ($M = 0.65$, $SD = 0.13$) relative to groups not engaging in pre-discussion small talk ($M = 0.69$, $SD = 0.05$). The causal mediation analysis revealed a significant negative indirect effect of small talk on entitativity via synchrony ($B = -.08$, 95%-CI [-0.17, -0.01], suggesting that the relationship between small talk and entitativity is partially mediated by synchrony

The preliminary findings of our study illuminate the complex relationship between small talk, synchrony, and entitativity within groups, suggesting that synchrony plays a nuanced role in the relationship between small talk and group entitativity. The presence of a significant direct effect alongside a significant indirect effect suggests that while small talk directly enhances group entitativity, its effect on reducing synchrony indirectly diminishes entitativity to some extent. This complexity suggests that the role of small talk in group dynamics may not be straightforward and could involve additional mediating factors not examined in this study. Another potential explanation for the negative impact of small talk on synchrony might be that small talk diverts group members' focus from the discussion topics, potentially delaying their synchronization. Thus, in further analyses we plan to account for the temporal dynamics of synchrony, specifically investigating the influence of small talk on the rate at which groups achieve synchronization.

Conclusion

Our study underscores the intricate relationship between small talk, synchrony, and group entitativity, offering novel insights into group dynamics. Small talk directly increases group cohesion, yet its nuanced interplay with synchrony suggests a multifaceted influence on entitativity. These insights pave the way for further exploration into the mechanisms underpinning group entitativity.

Key References

Allen, J. A., Lehmann-Willenbrock, N., & Landowski, N. (2014). Linking pre-meeting communication to meeting effectiveness. *Journal of Managerial Psychology*, *29*(8), 1064–1081. https://doi.org/10.1108/JMP-09-2012-0265

Blanchard, A. L. (2021). The effects of COVID-19 on virtual working within online groups. *Group Processes & Intergroup Relations*, *24*(2), 290–296. https://doi.org/10.1177/1368430220983446

Blanchard, A. L., & Allen, J. A. (2023). The entitativity underlying meetings: Meetings as key in the lifecycle of effective workgroups. *Organizational Psychology Review*, *13*(4), 458–477. https://doi.org/10.1177/20413866221101341

Blanchard, A. L., Caudill, L. E., & Walker, L. S. (2020). Developing an entitativity measure and distinguishing it from antecedents and outcomes within online and face-to-face groups. *Group Processes & Intergroup Relations*, *23*(1), 91–108. https://doi.org/10.1177/1368430217743577

Blanchard, A. L., & McBride, A. (2020). Putting the "Group" in Group Meetings: Entitativity in Face-to-Face and Online Meetings. In A. L. Meinecke, J. A. Allen, & N. Lehmann-Willenbrock (Eds.), *Managing Meetings in Organizations* (Vol. 20, pp. 71–92). Emerald Publishing Limited. https://doi.org/10.1108/S1534-085620200000020004

Blanchard, A. L., McBride, A. G., & Allen, J. A. (2022). Perceiving meetings as groups: How entitativity links meeting characteristics to meeting success. *Psychology of Leaders and Leadership*, *25*(2), 90–113. https://doi.org/10.1037/mgr0000124

Bleakley, A., Rough, D., Edwards, J., Doyle, P., Dumbleton, O., Clark, L., Rintel, S., Wade, V., & Cowan, B. R. (2021). Bridging social distance during social distancing: Exploring social talk and remote collegiality in video conferencing. *Human–Computer Interaction*, 1–29. https://doi.org/10.1080/07370024.2021.1994859

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). Jossey-Bass/Wiley.

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science*, *3*(1), 14–25. https://doi.org/10.1002/bs.3830030103

Collins, R. (2005). *Interaction ritual chains*. Princeton University Press. https://press.princeton.edu/books/paperback/9780691123899/interaction-ritual-chains

Coupland, J., Coupland, N., & Robinson, J. D. (1992). "How are you?": Negotiating phatic communion. *Language in Society*, *21*(2), 207–230. https://doi.org/10.1017/S0047404500015268

Holmes, J. (2003). Small talk at work: Potential problems for workers with an intellectual disability. *Research on Language & Social Interaction*, *36*(1), 65–84. https://doi.org/10.1207/S15327973RLSI3601_4

Holmes, J., & Marra, M. (2004). Relational practice in the workplace: Women's talk or gendered discourse? *Language in Society*, *33*(03). https://doi.org/10.1017/S0047404504043039

Jackson, J. C., Jong, J., Bilkey, D., Whitehouse, H., Zollmann, S., McNaughton, C., & Halberstadt, J. (2018). Synchrony and Physiological Arousal Increase Cohesion and

Cooperation in Large Naturalistic Groups. *Scientific Reports*, *8*(1), Article 1. https://doi.org/10.1038/s41598-017-18023-4

Keyton, J., Caputo, J. M., Ford, E. A., Fu, R., Leibowitz, S. A., Liu, T., Polasik, S. S., Ghosh, P., & Wu, C. (2013). Investigating Verbal Workplace Communication Behaviors. *The Journal of Business Communication (1973)*, *50*(2), 152–169. https://doi.org/10.1177/0021943612474990

Koch, T., & Denner, N. (2022). Informal communication in organizations: Work time wasted at the water-cooler or crucial exchange among co-workers? *Corporate Communications: An International Journal*. https://doi.org/10.1108/CCIJ-08-2021-0087

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 Questions About Interrater Reliability and Interrater Agreement. *Organizational Research Methods*, *11*(4), 815–852. https://doi.org/10.1177/1094428106296642

Lehmann-Willenbrock, N., & Hung, H. (2023). A Multimodal Social Signal Processing Approach to Team Interactions. *Organizational Research Methods*, 10944281231202741. https://doi.org/10.1177/10944281231202741

Methot, J. R., Rosado-Solomon, E. H., Downes, P. E., & Gabriel, A. S. (2021). Office chitchat as a social ritual: The uplifting yet distracting effects of daily small talk at work. *Academy of Management Journal*, *64*(5), Article 5. https://doi.org/10.5465/amj.2018.1474

Mirivel, J. C., & Tracy, K. (2005). Premeeting Talk: An Organizationally Crucial Form of Talk. *Research on Language & Social Interaction*, *38*(1), 1–34. https://doi.org/10.1207/s15327973rlsi3801_1

Molinsky, A. (2013, February 27). The Big Challenge of American Small Talk. *Harvard Business Review*. https://hbr.org/2013/02/the-big-challenge-with-america

Nanninga, M. C., Zhang, Y., Lehmann-Willenbrock, N., Szlávik, Z., & Hung, H. (2017). Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 206–215. https://doi.org/10.1145/3136755.3136811

R Core Team. (2022). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Van Swol, L. M., & Kane, A. A. (2019). Language and Group Processes: An Integrative, Interdisciplinary Review. *Small Group Research*, *50*(1), 3–38. https://doi.org/10.1177/1046496418785019

Varni, G., Avril, M., Usta, A., & Chetouani, M. (2015). SyncPy: A Unified Open-source Analytic Library for Synchrony. *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*, 41–47. https://doi.org/10.1145/2823513.2823520

Whittaker, S., Frohlich, D., & Daly-Jones, O. (1994). Informal workplace communication: What is it like and how might we support it? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Celebrating Interdependence - CHI '94*, 131–137. https://doi.org/10.1145/191666.191726

Zhao, X., Lynch, J. G., Jr., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis. *Journal of Consumer Research*, *37*(2), 197–206. https://doi.org/10.1086/651257

B.2 Group Synchrony and Attitude [P8]

Abstract

A heating Earth often sparks heated discussions – especially concerning strategies to address the climate crisis. While group discussions (e.g., among colleagues) can foster shared attitudes, they may also amplify differences and lead a divergence in attitudes. Yet, fostering collective attitudes is vital for effective collaboration. In this study, we investigate to what extent acoustic-prosodic synchrony between group members in a sustainability discussion influences their attitudes towards the climate crisis. We video-recorded 30 groups of three students discussing potential decarbonization measures for their university and assessed individual attitudes before and after the discussion as endorsement of the new ecological paradigm. Group variability in attitude was quantified using the standard deviation per group. To measure acoustic-prosodic synchrony, we used the audio signals to calculate global convergence between group members. Preliminary findings suggest that paraverbal synchrony may play a role in shaping collective attitudes toward the climate crisis. Our interdisciplinary research enhances our understanding of how communication dynamics may influence environmental perceptions, shedding light on the potential for linguistic synchronization to impact shared beliefs about the climate crisis. This might be relevant for creating shared climate beliefs in different group settings, among others in organizational contexts.

Reference

V. Begemann and C.S. Hemshorn de Sanchez and N. Raj Prabhu and N. Lehmann-Willenbrock, "In sync with sustainability: Acoustic-prosodic synchrony and attitudes toward the climate crisis in group discussions", *53rd Deutsche Gesellschaft für Psychologie (DGP) Congress*, Berlin, Germany, September, 2024.

Copyright Notice

Authors' Contributions

This work was a collaboration between the Organizational Psychology group and the Signal Processing group, at the University of Hamburg. Vanessa Begemann and Clara S. Hemshorn de Sanchez, as organizational psychology experts, were the lead contributors, responsible for setting up the experiment, collecting the group interaction data, and conducting the analysis and experimental validation. Navin Raj Prabhu, as the signal processing expert, developed the algorithms for extracting synchrony- and mimicry-based features and provided these for subsequent analysis. Nale Lehmann-Willenbrock offered valuable theoretical feedback and participated in the manuscript review.

In Sync with Sustainability: Acoustic-Prosodic Synchrony and Attitudes toward the Climate Crisis in Group Discussions

Vanessa Begemann, Clara Hemshorn de Sánchez, Navin Raj Prabhu & Nale Lehmann-Willenbrock
Universität Hamburg

Abstract:
A heating Earth often sparks heated discussions – especially concerning strategies to address the climate crisis. While group discussions (e.g., among colleagues) can foster shared attitudes, they may also amplify differences and lead a divergence in attitudes. Yet, fostering collective attitudes is vital for effective collaboration. In this study, we investigate to what extent acoustic-prosodic synchrony between group members in a sustainability discussion influences their attitudes towards the climate crisis. We video-recorded 30 groups of three students discussing potential decarbonization measures for their university and assessed individual attitudes before and after the discussion as endorsement of the new ecological paradigm. Group variability in attitude was quantified using the standard deviation per group. To measure acoustic-prosodic synchrony, we used the audio signals to calculate global convergence between group members. Controlling for pre-discussion group variability in attitude, preliminary regression analysis results indicate a negative effect of global convergence on post-discussion group variability in attitude toward the climate crisis, (B = .04, SE = .03, p = .059), suggesting an increase in attitude convergence in groups. These preliminary findings suggest that paraverbal synchrony may play a role in shaping collective attitudes toward the climate crisis. Our interdisciplinary research enhances our understanding of how communication dynamics may influence environmental perceptions, shedding light on the potential for linguistic synchronization to impact shared beliefs about the climate crisis. This might be relevant for creating shared climate beliefs in different group settings, among others in organizational contexts.