Model-Based Deep Speech Enhancement for Improved Interpretability and Robustness

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

von

Huajian Fang

Hamburg, 2025

Huajian Fang: *Model-Based Deep Speech Enhancement for Improved Interpretability and Robustness*

DISSERTATION COMMITTEE:

Prof. Dr.-Ing. Timo Gerkmann (supervisor and reviewer)

Prof. Dr. Stefan Wermter (supervisor)

Prof. Dr.-Ing. Dorothea Kolossa (reviewer)

Prof. Dr. Chris Biemann (chair)

Prof. Dr. Jianwei Zhang (deputy chair)

DATE OF SUBMISSION:

12.02.2025

DATE OF DISPUTATION:

02.06.2025

Abstract

Technology advancements profoundly impact numerous aspects of life, including how we communicate and interact. For instance, hearing aids enable hearing-impaired or elderly people to participate comfortably in daily conversations; telecommunications equipment lifts distance constraints, enabling people to communicate remotely; smart machines are developed to interact with humans by understanding and responding to their instructions. These applications involve speech-based interaction not only between humans but also between humans and machines. However, the microphones mounted on these technical devices can capture both target speech and interfering sounds, posing challenges to the reliability of speech communication in noisy environments. For example, distorted speech signals may reduce communication fluency among participants during teleconferencing. Additionally, noise interference can negatively affect the speech recognition and understanding modules of a voice-controlled machine. This calls for speech enhancement algorithms to extract clean speech and suppress undesired interfering signals, improving the overall quality and intelligibility of speech.

Traditional speech enhancement algorithms often rely on simplifying assumptions, such as slowly changing noise, to estimate the parameters required for clean speech estimators. This may lead to less than satisfactory results in acoustically challenging scenarios. In recent years, the field has seen great strides through deep learning-based algorithms. The success of deep learning stems largely from its universal function approximation capability and scalability to large datasets. In particular, deep predictive approaches have received widespread attention due to their remarkable flexibility in incorporating key features of the target speech into various stages of the speech enhancement framework. These stages include input feature processing, network architecture design, training objective formulation, and optimization strategy development. Essentially, deep predictive methods aim to learn a mapping between noisy mixtures and clean speech by training deep neural networks (DNNs) on a large number of paired noisy-clean speech samples. However, the performance of these algorithms depends heavily on the quantity and diversity of training data. As a result, performance degradation often occurs when there is a data mismatch between training and testing, known as the generalization problem. Moreover, predictive approaches are typically framed as problems with a single output, which may result in erroneous estimates for complex and unseen samples without any indication of uncertainty. Indeed, due to the black-box nature of DNNs, deep learning-based algorithms produce clean speech estimates in a non-transparent manner, making them difficult to interpret. In this thesis, we aim to incorporate statistical models into DNN-based speech enhancement to improve its robustness and interpretability.

The first part of the thesis explores these ideas from the perspective of uncertainty. We augment predictive speech enhancement with an uncertainty estimation task, such that the network model can provide not only clean speech estimates but also their associated predictive uncertainty. Furthermore, since generic Bayesian methods for uncertainty modeling in deep learning usually involve costly sampling processes, this thesis seeks to leverage statistical knowledge from the speech processing domain to efficiently estimate uncertainty with minimal computational overhead. We experimentally demonstrate that the proposed uncertainty-augmented framework effectively identifies when predictions deviate significantly from the true data by producing large uncertainty estimates. This allows us to assess the model's confidence in predictions when clean speech ground truth is unavailable. Additionally, we show that the uncertainty-augmented methods grounded in statistical modeling improve speech enhancement performance compared to methods that predict a single filter mask only. Next, we explore the direct use of uncertainty estimates for speech enhancement tasks. This includes unsupervised domain adaptation, where we utilize uncertainty-based filtering to select high-quality pseudo-targets to alleviate generalization issues. In another application, alongside audio inputs, we further explore modeling uncertainty originating from distorted video signals in an audio-visual phoneme classification task and demonstrate how to exploit modality-wise uncertainty to achieve more effective and robust multimodal fusion.

In the second part of the thesis, we investigate the issues of interpretability and robustness by focusing on deep generative approaches. In contrast to predictive approaches that learn a deterministic mapping between noisy and clean speech, deep generative approaches aim to learn prior distributions of given data and reuse this knowledge to perform speech enhancement during inference. In the thesis, we consider a specific group of methods, which use a variational autoencoder (VAE) to learn a prior distribution of clean speech and combine it with an untrained non-negative matrix factorization (NMF)-based noise model to estimate a filter mask for speech enhancement. The statistically interpretable VAE-NMF framework exhibits an improved generalization ability to unseen acoustic conditions compared to predictive methods. However, training the VAE solely with clean speech makes it susceptible to noise interference during testing, especially for inputs with low signal-to-noise ratios. In this part, we aim to improve overall robustness in difficult acoustic conditions by augmenting separately the speech and noise models with noise information. The resulting noise-aware speech and noise models retain the high interpretability provided by statistical modeling while at the same time exhibiting improved speech enhancement performance in acoustically challenging environments.

Zusammenfassung

Der technologische Fortschritt hat tiefgreifende Auswirkungen auf zahlreiche Aspekte des Lebens, einschließlich darauf, wie wir kommunizieren und interagieren. Beispielsweise ermöglichen Hörgeräte es hörgeschädigten oder älteren Menschen, bequem an alltäglichen Gesprächen teilzunehmen; Telekommunikationsgeräte heben Entfernungsbeschränkungen auf und ermöglichen es Menschen, aus der Ferne miteinander zu kommunizieren; intelligente Maschinen werden entwickelt, um mit Menschen zu interagieren, indem sie deren Anweisungen verstehen und darauf reagieren. Diese Anwendungen beinhalten sprachbasierte Interaktionen nicht nur zwischen Menschen, sondern auch zwischen Menschen und Maschinen. An solchen technischen Geräten angebrachte Mikrofone können jedoch sowohl die Sprache der Zielquelle als auch Störgeräusche erfassen, was die Zuverlässigkeit der Sprachkommunikation in lauten Umgebungen beeinträchtigt. So können beispielsweise verzerrte Sprachsignale den Kommunikationsfluss zwischen den Teilnehmern einer Telefonkonferenz beeinträchtigen. Außerdem können Störgeräusche die Module einer sprachgesteuerten Maschine zur Spracherkennung und zum Sprachverstehen negativ beeinflussen. Daher sind Algorithmen zur Sprachverbesserung erforderlich, die das saubere Sprachsignal extrahieren und unerwünschte Störsignale unterdrücken, um die Gesamtqualität und Verständlichkeit der Sprache zu verbessern.

Herkömmliche Algorithmen zur Sprachverbesserung stützen sich häufig auf vereinfachende Annahmen, wie z. B. sich nur langsam verändernde Störgeräusche, um die erforderlichen Parameter für die Schätzung der sauberen Sprache zu bestimmen. Dies kann in akustisch schwierigen Szenarien zu unbefriedigenden Ergebnissen führen. In den letzten Jahren wurden auf diesem Gebiet durch Algorithmen, die auf Deep Learning basieren, große Fortschritte erzielt. Der Erfolg von Deep Learning beruht zum großen Teil auf seiner universellen Fähigkeit zur Approximation mathematischer Funktionen und seiner Skalierbarkeit für große Datensätze. Insbesondere prädiktive Deep-Learning-Ansätze haben aufgrund ihrer bemerkenswerten Flexibilität bei der Einbeziehung von Kernmerkmalen des Zielsprachsignals in verschiedenen Schritten des Sprachverbesserungsalgorithmus große Beachtung gefunden. Zu diesen Schritten gehören die Verarbeitung von Merkmalen des Eingangssignales, das Design der Netzwerkarchitektur, die Formulierung von Trainingszielen und die Entwicklung von Optimierungsstrategien. Im Wesentlichen zielen prädiktive Deep-Learning-Methoden darauf ab, eine Abbildung zwischen verrauschten Mischsignalen und sauberer Sprache zu erlernen, wobei Deep Neural Networks (DNNs) auf einer großen Anzahl von gepaarten verrauschten und sauberen Sprachbeispielen trainiert werden. Die Leistung dieser Algorithmen hängt jedoch stark von der Menge und Vielfalt der Trainingsdaten ab. Daher kommt es häufig zu einer Leistungsverschlechterung,

wenn die Daten zwischen Training und Test nicht übereinstimmen, was als Generalisierungsproblem bekannt ist. Darüber hinaus werden prädiktive Ansätze in der Regel als Probleme mit einer einzigen möglichen Lösung betrachtet, was zu fehlerhaften Schätzungen für komplexe und ungesehene Beispiele ohne jegliche Angabe von Unsicherheiten führen kann. Aufgrund des Black-Box-Charakters von DNNs produzieren Deep-Learning-Algorithmen saubere Sprachschätzungen auf eine nicht interpretierbare Weise. In dieser Arbeit zielen wir darauf ab, statistische Modelle in die DNN-basierte Sprachverbesserung zu integrieren, um ihre Robustheit und Interpretierbarkeit zu verbessern.

Im ersten Teil der Arbeit werden diese Ideen aus einer auf Unsicherheit fokussierten Perspektive untersucht. Wir erweitern die prädiktive Sprachverbesserung darum, zusätzlich Unsicherheiten zu schätzen, sodass das Netzwerkmodell nicht nur saubere Sprachschätzungen, sondern auch die damit verbundene prädiktive Unsicherheit liefern kann. Da generische Bayes'sche Methoden zur Unsicherheitsmodellierung im Deep Learning in der Regel kostspielige Sampling-Prozesse involvieren, wird in dieser Arbeit versucht, statistisches Wissen aus dem Bereich der Sprachverarbeitung zu nutzen und so die Unsicherheit unter minimalem zusätzlichem Rechenaufwand effizient zu schätzen. Wir demonstrieren experimentell, dass das vorgeschlagene um Unsicherheitsschätzung erweiterte Rahmenwerk effektiv Fälle identifiziert, in denen die Vorhersagen signifikant von den wahren Daten abweichen, indem es dort große Unsicherheiten ausgibt. Dies ermöglicht es uns, das Vertrauen des Modells in seine eigenen Vorhersagen zu bewerten, auch wenn das saubere Referenzsprachsignal nicht verfügbar ist. Darüber hinaus zeigen wir – im Vergleich zu Methoden, welche nur eine einzelne Filtermaske schätzen – dass unsere auf statistischer Modellierung basierenden um Unsicherheitsschätzung erweiterten Methoden die Qualität der Sprachverbesserung verbessern. Als nächstes untersuchen wir die direkte Nutzung der geschätzten Unsicherheiten für Aufgaben der Sprachverbesserung. Dazu gehört die sogenannte unsupervised domain adaptation, bei der wir eine auf Unsicherheit basierende Filterung nutzen, um qualitativ hochwertige Pseudo-Ziele auszuwählen und damit Generalisierungsprobleme abzumildern. In einer weiteren Anwendung für die audiovisuelle Klassifikation von Phonemen erforschen wir neben Audio-Eingabesignalen auch die Modellierung der Unsicherheiten von verzerrten Videosignalen, und zeigen, wie man modalitätsbezogene Unsicherheiten nutzen kann, um eine effektivere und robustere Fusion dieser multimodalen Signale zu erreichen.

Im zweiten Teil der Arbeit untersuchen wir die Fragen der Interpretierbarkeit und Robustheit, und fokussieren uns hierfür auf generative Ansätze des Deep Learning. Im Gegensatz zu prädiktiven Ansätzen, die eine deterministische Abbildung von verrauschter zu sauberer Sprache erlernen, zielen generative Deep-Learning-Ansätze darauf ab, a-priori-Verteilungen der gegebenen Daten zu erlernen und dieses Wissen wiederzuverwenden, um eine Sprachverbesserung durchzuführen. In dieser Arbeit betrachten wir eine spezielle Klasse von Methoden zur Schätzung einer Filtermaske für die Sprachverbesserung, welche einen variational autoencoder (VAE) verwenden, um eine a-priori-Verteilung von sauberer Sprache zu lernen und diese mit einem untrainierten, auf einer non-negative matrix factorization (NMF) basierenden Rauschmodell kombinieren. Dieses statistisch interpretierbare VAE-NMF-Framework zeigt im Vergleich zu prädiktiven Methoden eine verbesserte Generalisierungsfähigkeit für zuvor ungesehene akustische Szenarien. Wird der VAE jedoch auss-

chließlich mit sauberer Sprache trainiert, ist er anfällig für Störgeräusche beim Testen, insbesondere bei Eingaben mit niedrigem Signal-Rausch-Verhältnis. In diesem Teil der Arbeit versuchen wir, die allgemeine Robustheit unter schwierigen akustischen Bedingungen zu verbessern, indem wir die Sprach- und Störsignalmodelle separat mit Informationen über das Störsignal augmentieren. Die daraus resultierenden Sprach- und Störsignalmodelle erhalten die hohe Interpretierbarkeit der statistischen Modellierung und zeigen gleichzeitig eine Verbesserung der Sprache in akustisch schwierigen Szenarien.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Prof. Dr.-Ing. Timo Gerkmann and Prof. Dr. Stefan Wermter. I would like to thank Prof. Dr.-Ing. Timo Gerkmann for his patience, encouragement, constant support, and insightful guidance throughout my PhD journey. Thank you for always being there when I needed to discuss problems and for helping me see things from new perspectives whenever I felt stuck. I'm also grateful to Prof. Dr. Stefan Wermter for his valuable support and constructive feedback. Our research discussions, especially at the beginning of my PhD, helped guide me onto the right path. Thank you for your inspiring advice whenever I approached you — your input always helped me reflect and refine my ideas.

I would like to thank Prof. Dr.-Ing. Dorothea Kolossa for reviewing my thesis, Prof. Dr. Chris Biemann for chairing the disputation committee, and Prof. Dr. Jianwei Zhang for serving as the deputy chair.

Working in two groups was another blessing as it meant I had the support of colleagues on both sides, doubling the fun, learning, and great memories we shared. I would like to thank all my colleagues from the Signal Processing Group and the Knowledge Technology Group. Working with you has made my journey so much fun. I learned a lot from our research chats, group meetings, and casual whiteboard brainstorming. The conference trips, lunches, coffee breaks, and after-work get-togethers are memories I will always cherish.

I want to thank my parents, my sister Lily, and my partner Shuwen for their endless support and encouragement and for always believing in me.

List of Acronyms

CGMM complex Gaussian mixture model

CNN convolutional neural network

DNN deep neural network

ELBO evidence lower bound

EM expectation-maximization

GRU gated recurrent unit

IS Itakura-Saito

KL Kullback-Leibler

LSTM long short-term memory

MAE mean absolute error

MAP maximum a posterior

MC Monte Carlo

MCEM Monte Carlo expectation maximization

MCMC Markov chain Monte Carlo

MH Metropolis-Hastings

MMSE minimum mean square error

MSE mean squared error

MVDR minimum variance distortionless response

NMF non-negative matrix factorization

PSD power spectrum density

RNN recurrent neural network

SDR signal-to-distortion ratio

SI-SDR scale-invariant signal-to-distortion ratio

SNR signal-to-noise ratio

STFT short-time Fourier transform

VAE variational autoencoder

Contents

CHAPTER 1

Introduction

1.1 Motivation

Speech communication plays a vital role in human interaction; it is an intuitive way to connect with people and express emotions. Nowadays, speech communication is no longer limited by distance, but can also take place remotely through teleconferencing equipment. Furthermore, the rapid development of technology has resulted in the emergence of many intelligent machines, such as smart-home devices, autonomous vehicles, and humanoid interactive robots. These smart systems can sense and interact with the surrounding environment through the key modality of acoustic signals. Thus, speech-based interaction occurs not only between humans but also between humans and machines. While speech communication devices and smart machines have become ubiquitous, they face the same challenge that the target speech recorded by microphones is inevitably distorted by interfering noise. Communication becomes more challenging and less comprehensible when noise interference becomes severe. The same holds for interaction with machines, where heavily distorted target signals pose great challenges to the speech recognition and understanding modules of smart devices. This calls for robust speech enhancement algorithms, which extract the target speech by suppressing undesired interference signals.

Common environmental interference signals encompass a wide variety of background noises, ranging from traffic noise recorded on the street to human-generated noise from indoor activities. The nature of environmental noise is closely related to the application scenario, exhibiting diverse spectral characteristics and varying input signal-to-noise ratios (SNRs). Notably, ensuring the robustness of speech enhancement algorithms across various acoustic conditions is crucial yet challenging. Besides background noise disturbances, human-robot interaction introduces additional self-created noise, commonly referred to as ego-noise [1], mainly caused by the internal motors and mechanical parts of robots. Ego-noise generated by motors moving at different speeds is non-stationary, and its time-frequency analysis exhibits broadband characteristics. This implies that noise interference can largely overlap with the target speech in the spectral domain. Moreover, for small-sized robots like NAO [2, 3], microphones are positioned close to joints. This proximity to noise sources often leads

to low SNR recordings. These observations make ego-noise reduction also a challenging problem. However, compared to environmental noise that varies with the surrounding environment, ego-noise often exhibits less diversity due to the limited degrees of freedom of motors' motion. Therefore, effective ego-noise reduction methods can be developed by leveraging the relatively less diverse nature of ego-noise [1]. In contrast, speech enhancement algorithms handling environmental noise often target high generalization ability so that one algorithm can be applied to various acoustic scenarios [4, 5].

Various speech enhancement algorithms have been proposed to suppress environmental noise or ego-noise, achieving different degrees of success. Traditional approaches estimate the statistical parameters of speech and noise from noisy mixtures using, e.g., power spectrum density (PSD) tracking techniques based on voice activity or speech presence probability [6, 7]. However, it is often assumed that noise is changing more slowly than speech, which leads to limited enhancement performance when speech is distorted by non-stationary noise. In contrast, machine learning-based methods do not necessarily need to follow these simplifying assumptions. They can optionally learn prior knowledge from data and often yield superior performance. For example, dictionary-based algorithms [8–10] can be used to learn the temporal-spectral patterns in the time-frequency domain from training data, providing a good fit for modeling structured signals. As an illustration, each phoneme in a language exhibits a specific spectral pattern and a group of phonemes shows additional temporal characteristics. Besides these observed in speech, we may also detect certain spectral features in noise signals [1]. Further performance gains can be obtained through deep learning-based methods, which have received widespread attention due to their excellent ability to learn meaningful and complex representations from data.

Deep neural networks (DNNs) can arbitrarily approximate any functions, offering great flexibility to integrate key characteristics of the target signal into the algorithm development. The versatility has facilitated widespread use in the field of speech enhancement, from replacing building blocks of conventional approaches to developing end-to-end speech enhancement frameworks. These methods are primarily data-driven and have demonstrated remarkable performance improvements compared to traditional methods. In particular, *deep predictive approaches* have emerged as the dominant technique in DNN-based speech enhancement (also called *deep speech enhancement* interchangeably in the thesis). In machine learning, the problem of learning a function that maps inputs to outputs from a paired dataset can be modeled using a conditional probability function. When the output variable is discrete, it is typically referred to as a *discriminative model*, which is intuitively understood as discriminating between the classification boundaries among different classes. In contrast, it is referred to as a *regression model* when the output variable is continuous. In this thesis, we follow the discussion in [11] and use a unified term *predictive model* to cover both the regression and classification settings. For instance, prevailing predictive masking approaches learn to estimate multiplicative filter masks and extract clean speech by applying these masks to the corresponding noisy mixtures. The training process is guided by the noisy-to-clean mapping relationships established through labeled training data. However, the performance of these algorithms is closely tied to the

diversity and quantity of training data. Their robustness is not guaranteed under acoustic conditions not covered by training data, leading to generalization issues. Various aspects of these approaches have evolved significantly to overcome the limitations, including the development of informative input features [4, 12], advanced network architectures [13–16], effective loss functions [17], and improved optimization schemes [18, 19]. Nevertheless, gaining robustness under complex and unseen test conditions remains an ongoing research topic. Furthermore, despite being widely used, DNNs are typically treated as tools of a black-box nature and provide clean speech estimates in a non-transparent manner, making the network model's estimation behavior difficult to interpret. This thesis investigates the challenges of generalization and interpretability from two perspectives:

- We want the predictive model to provide a reliability indicator for its predictions through uncertainty modeling.
- Instead of relying exclusively on the deterministic mapping function learned from labeled data, we want to explore statistically interpretable deep generative models that can learn data distributions.

The thesis revolves around incorporating statistical modeling into deep speech enhancement, showing how these ideas are realized through this incorporation. Moreover, this enables the combination of the regularization benefits of statistical modeling with the non-linear modeling capabilities of DNNs. The thesis mainly contains two parts. The **first** part is based on prevailing deep predictive masking approaches. The inevitable data mismatch between training and testing raises the question: Can we enable neural networks to provide confidence in their predictions? Therefore, our focus in the first part is to investigate uncertainty modeling in the context of DNN-based speech enhancement. Uncertainty suggests discrepancies between predictions and the true data, and uncertainty quantification should be considered an important feature of deep speech enhancement algorithms, in addition to achieving high performance. Thus, the questions of how to model uncertainty in the context of deep speech enhancement, how to incorporate domain-specific statistical knowledge to facilitate efficient uncertainty modeling, and how to employ uncertainty arising from DNN-based methods for further use are of great interest. In the **second** part, we study the interpretability and generalization issues by focusing on deep generative approaches, more specifically, a framework that integrates a variational autoencoder (VAE)-based speech model and non-negative matrix factorization (NMF)-based noise model [20]. It is observed that the performance of deep predictive approaches depends heavily on paired training data, leading to challenges in generalizing well to out-of-distribution samples during testing. These may include unseen noise types, SNRs, speakers, sound loudness, and acoustic properties. In contrast, the VAE-NMF framework learns a prior distribution of clean speech and reuses this knowledge for speech enhancement in a statistically principled manner. This exhibits improved generalization capabilities over comparable predictive baselines [21, 22]. However, it is difficult to ensure that such a method can consistently present good performance in adverse acoustic situations, including unseen speakers, non-stationary noise distortions, and low SNRs. Building upon this framework grounded in statistical modeling, the second part of the thesis focuses on how prior noise information can be used to refine

the speech and noise models separately in order to improve the overall robustness of the algorithm in challenging acoustic environments.

In the rest of this chapter, we first describe statistical assumptions relevant to this thesis in Section 1.2. In Section 1.3, we provide an overview of deep predictive speech enhancement, followed by uncertainty modeling in predictive learning. We introduce deep generative approaches in Section 1.4. We present the thesis structure and an overview of the related publications in Section 1.5 and 1.6 respectively. We categorize our publications into two groups based on the research topics. Chapter 2 contains the accepted versions of articles focusing on uncertainty modeling. Chapter 3 includes the accepted versions of articles exploring deep generative approaches. Finally, we summarize our findings and discuss potential future research in Chapter 4.

1.2 Statistical Modeling in Speech Enhancement

In this section, we focus on the speech enhancement task, which aims to remove noise interference from noisy mixtures to improve the quality and intelligibility of target speech. The majority of traditional speech enhancement algorithms are formulated in the time-frequency domain, rather than directly operating in the time domain [5]. This is motivated by the fact that time-frequency representations of signals often show distinctive structures. For example, it can be observed that vowel sounds of speech in the time-frequency domain exhibit harmonic structures, which are integer multiples of the fundamental frequency. Moreover, since speech utterances are composed of a sequence of phonemes, a combination of multiple phonemes may exhibit specific temporal structures, which can be leveraged to differentiate target speech from other sources effectively. Not only speech but also some noise sources exhibit specific temporal-spectral structures. For instance, environmental noise recorded on streets may contain, e.g., bus engine noise, which is relatively stationary and may slowly change over time. Spectral representations of ego-noise generated by the robot's motors and mechanical parts also exhibit harmonic structures [1]. These observations make signal modeling in the time-frequency domain more efficient and effective than direct processing in the time domain. Moreover, sound sources often display sparse characteristics in the time-frequency domain, where only a limited number of time-frequency bins contain large amplitude values while the rest remain relatively small. The sparsity leads to less overlap of these large amplitude values in the time-frequency domain [23]. Therefore, recognizing and distinguishing these unique temporal-spectral characteristics is crucial in effectively restoring clean speech and suppressing undesirable noise.

1.2.1 Clean Speech Estimators

A widely used tool to convert time-domain signals to their time-frequency representations is the short-time Fourier transform (STFT). Its widespread adoption can be attributed to several desirable properties. One of the advantages is the computational efficiency achieved through the fast Fourier transform. Additionally, the Fourier basis functions are orthogonal and can easily decompose a

windowed signal into its frequency components. We can perfectly reconstruct the original signal from its unmodified STFT spectrogram with properly designed transform parameters. Furthermore, the time-frequency representations provide information about both the time and frequency content, and this provides a good fit for the spectral analysis of non-stationary signals, whose characteristics change over time. The window length is often chosen to reach a desired trade-off between time and frequency resolution. In STFT-based speech enhancement methods, it is generally assumed that clean speech is distorted by additive and independent noise, which can be mathematically defined as:

$$X_{ft} = S_{ft} + N_{ft}, \tag{1.1}$$

where X_{ft}, S_{ft}, and N_{ft} denote spectral coefficients of the noisy mixture, clean speech, and additive noise, respectively. $f \in \{1, \cdots, F\}$ and $t \in \{1, \cdots, T\}$ are the frequency index and time frame index, respectively. In the time-frequency domain, speech enhancement can be achieved by applying a multiplicative filter W_{ft} to the noisy mixture, giving a clean speech estimate $\widehat{S}_{ft}$:

$$\widehat{S}_{ft} = W_{ft} X_{ft}. \tag{1.2}$$

Bayesian modeling considers these spectral coefficients as realizations of random variables and provides a principled method to obtain filter masks. Typically, the speech-plus-noise model is applied independently to each time-frequency bin to facilitate mathematical computation and derive an analytical solution for the clean speech estimator in a Bayesian optimal sense [5]. Various Bayesian estimators have been developed based on different statistical assumptions about speech and noise, aiming to restore either the spectral coefficients of the STFT or the spectral magnitudes. Common statistical estimators include the maximum a posterior (MAP) estimator, which finds the mode of the posterior distribution of clean speech, and the minimum mean square error (MMSE) estimator, which minimizes the average squared error of the clean speech estimate.

Different models assume different statistical beliefs about the signal of interest, referred to as the *prior*. Statistical magnitude estimators can be derived by combining a magnitude prior (e.g., generalized-Gamma distribution) with the uniform-distributed phase assumption [24], while estimators of the complex spectral coefficients are often derived based on the independent assumption made between the real and imaginary parts of the complex spectral coefficients [5]. For example, assuming that speech is degraded by additive noise and both follow a circularly symmetric complex Gaussian distribution yields the well-known *Wiener filter* [25]. Formally, the speech and noise priors are defined as follows:

$$S_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2_{S,ft}) \quad \text{and} \quad N_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2_{N,ft}), \tag{1.3}$$

where $\mathcal{N}_{\mathbb{C}}$ denotes the complex Gaussian distributions. $\sigma^2_{S,ft}$ and $\sigma^2_{N,ft}$ are the speech and noise variances, respectively. Given the additive and independence assumptions, the likelihood $p(X_{ft}|S_{ft})$ is thus given in the form of complex Gaussian with mean S_{ft} and variance $\sigma^2_{N,ft}$ and the evidence $p(X_{ft})$ follows a zero-mean complex Gaussian with variance $\sigma^2_{S,ft} + \sigma^2_{N,ft}$. By applying Bayes' theorem

$p(S_{ft}|X_{ft}) = \frac{p(X_{ft}|S_{ft})p(S_{ft})}{p(X_{ft})}$, we can obtain the posterior distribution of clean speech [25]:

$$p(S_{ft}|X_{ft}) = \frac{1}{\pi\lambda_{ft}} \exp\left(-\frac{|S_{ft} - W_{ft}X_{ft}|^2}{\lambda_{ft}}\right), \tag{1.4}$$

where $W_{ft} = \frac{\sigma_{S,ft}^2}{\sigma_{S,ft}^2+\sigma_{N,ft}^2}$ is referred to as Wiener filter and $\lambda_{ft} = \frac{\sigma_{S,ft}^2\sigma_{N,ft}^2}{\sigma_{S,ft}^2+\sigma_{N,ft}^2}$ is the variance of the posterior distribution. Therefore, clean speech can be estimated by taking the mean of the posterior. Since the posterior is also normally distributed and thus unimodal and symmetric (skewness of zero), the MMSE estimator and the MAP estimator of spectral coefficients are equivalent, i.e., both are represented by the Wiener filter. The local Gaussian model has received widespread attention and has been combined with many variance modeling techniques to extract the target source in many applications [26, 27], partially due to its simplicity. While the speech and noise variances in the Wiener filter can be estimated using traditional statistical methods [7, 25, 6], its performance is often limited when handling non-stationary noise distortions. This problem can be largely alleviated when leveraging DNNs to estimate filter masks directly. Furthermore, supervised DNN-based approaches trained on synthetic noisy-clean speech pairs with the common mean squared error (MSE) cost function:

$$\mathcal{L} = \frac{1}{FT}\sum_{f,t}|S_{ft} - W_{ft}X_{ft}|^2 \tag{1.5}$$

can be interpreted as implicitly assuming the complex Gaussian model of speech and noise and a constant variance for all time-frequency bins [17], as will be illustrated in [P2].

Noise is often seen as a sum of numerous independent sources and is typically modeled by a complex Gaussian distribution, which is motivated by the central limit theorem [28, 5]. In contrast, the histogram of speech spectral coefficients has shown super-Gaussianity, i.e., the histograms of real/imaginary parts of the spectral coefficients are more peaky and heavy-tailed than Gaussian [5, 29]. This observation allows the use of a super-Gaussian prior to fit speech spectral coefficients [30, 29, 31]. However, super-Gaussian speech priors are often accompanied by complex mathematical derivations of MMSE solutions. The resulting clean speech estimators may require the accurate computation of complex functions, which is computationally demanding [29, 24]. To overcome the complexity, Astudillo [32] has proposed to approximate super-Gaussian priors by extending the complex Gaussian to the complex Gaussian mixture model (CGMM). The mixture model is generally referred to as a linear combination of basis distributions; the CGMM here takes as the basis distribution the complex Gaussian. The Gaussian mixture density model possesses the advantage of being able to approximate any continuous density function to arbitrary accuracy with a sufficient number of basis components [33], which provides a good fit for modeling super-Gaussian characteristics of the speech coefficients. Thus, we can model the spectral coefficients of speech and noise with a mixture of

zero-mean circularly symmetric complex Gaussians

$$S_{ft} \sim \sum_{i=1}^{I} \Omega(i)\, \mathcal{N}_{\mathbb{C}}(0,\, \sigma_{i,ft}^2), \quad N_{ft} \sim \sum_{j=1}^{J} \Omega(j)\, \mathcal{N}_{\mathbb{C}}(0,\, \sigma_{j,ft}^2)\,, \tag{1.6}$$

where $\sigma_{i,ft}^2$ represents the variance of the i-th complex Gaussian component of the speech CGMM, and $\sigma_{j,ft}^2$ denotes the variance of the j-th complex Gaussian component of the noise CGMM. Similarly, the likelihood $p(X_{ft}|S_{ft})$ is modeled by a CGMM centered at S_{ft}. With the prior and likelihood specified, we can apply Bayes' theorem to obtain the corresponding posterior of speech spectral coefficients:

$$p(S_{ft}|X_{ft}) = \sum_{i=1}^{I}\sum_{j=1}^{J} \Omega(i,j|X_{ft}) \frac{1}{\pi\lambda_{ij,ft}} \exp\left(-\frac{|S_{ft} - W_{ij,ft}X_{ft}|^2}{\lambda_{ij,ft}}\right), \tag{1.7}$$

where $W_{ij,ft} = \frac{\sigma_{i,ft}^2}{\sigma_{i,ft}^2+\sigma_{j,ft}^2}$ and $\lambda_{ij,ft} = \frac{\sigma_{i,ft}^2\sigma_{j,ft}^2}{\sigma_{i,ft}^2+\sigma_{j,ft}^2}$ are the Wiener filter and the posterior's variance of the mixture Gaussian pair (i,j), respectively. $\Omega(i,j|X_{ft})$ denotes the posterior's mixture weights, which sum to 1. As a result, each Gaussian component, indexed by ij, in the posterior provides a Wiener estimate:

$$\widehat{S}_{ij,ft} = W_{ij,ft}X_{ft} = \frac{\sigma_{i,ft}^2}{\sigma_{i,ft}^2 + \sigma_{j,ft}^2} X_{ft}. \tag{1.8}$$

A clean speech estimate can then be obtained by computing the expectation of the posterior CGMM, yielding:

$$\mathbb{E}(S_{ft}|X_{ft}) = \int S_{ft}\; p(S_{ft}|X_{ft}) dS_{ft} = \sum_{i=1}^{I}\sum_{j=1}^{J} \Omega(i,j|X_{ft})\widehat{S}_{ij,ft}\;. \tag{1.9}$$

With the extension to the mixture priors, one can derive a closed-form MMSE solution, allowing for more accurate modeling of the speech target in the time-frequency domain. Unlike the posterior mean, finding the MAP solution of the mixture model is challenging due to the inherent complexity of the multi-mode nature of the posterior. While the accurate estimation of the posterior mode may involve complex iterative optimization procedures, a simplified approximation can be determined as a practical alternative, i.e., by selecting the mode of a single Gaussian component of the CGMM posterior guided by the mixing coefficients $\Omega(i,j|X_{ft})$ [33]. However, this is not guaranteed to obtain a global mode.

Clean speech estimation based on statistical modeling of complex spectral coefficients provides a principled way to obtain clean speech filter masks. Nevertheless, the efficacy of statistical estimators relies heavily on accurate parameter estimation. DNNs have emerged as a universal model that possesses powerful non-linear modeling capabilities and allows the use for a variety of purposes, as will be discussed in Section 1.3 and 1.4. In this thesis, we will discuss integrating statistical models into deep speech enhancement frameworks in various settings. This integration aims to combine the universal approximation capabilities provided by neural networks with statistical modeling as a way

to improve the robustness of the algorithm as well as to provide more interpretability of the estimation behavior.

Multi-Channel Clean Speech Estimator

In the last section, we discussed the single-channel case, where speech enhancement algorithms extract clean speech by differentiating temporal-spectral features of speech from that of noise. Multi-channel speech enhancement algorithms can leverage additional spatial information, which characterizes the sound propagation path between the sources and microphones. Let $\mathtt{X}_{ft} \in \mathbb{C}^M$ denote the multi-channel noisy spectral coefficients and M is the number of microphones of a microphone array. Given the additive speech-plus-noise assumption, the spectral coefficients of the noisy mixture signal recorded by a microphone array can be decomposed into

$$\mathtt{X}_{ft} = \mathtt{S}_{ft} + \mathtt{N}_{ft}, \tag{1.10}$$

where $\mathtt{S}_{ft} \in \mathbb{C}^M$ and $\mathtt{N}_{ft} \in \mathbb{C}^M$ are the spectral coefficients of the spatial images of speech and noise. Similar to the single-channel case, a multiplicative filter can be applied to extract the clean speech target

$$\widehat{\mathtt{S}}_{ft} = \mathbf{W}_{ft}^H \mathtt{X}_{ft}, \tag{1.11}$$

where $\mathbf{W}_{ft} \in \mathbb{C}^{M \times M}$ denotes a multi-channel filter mask at f-th frequency bin and t-th time frame and $(\cdot)^H$ denotes the conjugate transpose operator. $\widehat{\mathtt{S}}_{ft}$ denotes an clean speech estimate at f-th frequency bin and t-th time frame . It can be observed that this filtering process extracts clean speech by performing a linear operation with respect to a local multi-channel time-frequency bin. A typical example of such a filter mask is the multi-channel Wiener filter, which can be derived under the MMSE criterion [34]:

$$\mathbf{W}_{ft} = \arg\min_{\mathbf{W}} \mathbb{E}\{||\mathbf{W}_{ft}^H \mathtt{X}_{ft} - \mathtt{S}_{ft}||_2^2\}. \tag{1.12}$$

It is known that the MMSE solution is equivalent to finding the mean of the speech posterior [25]. Thus, we can also obtain the multi-channel Wiener filter by deriving the multi-channel speech posterior. For this, we can model multi-channel speech and noise spectral coefficients using zero-mean multivariate Gaussian distributions [27]:

$$\mathtt{S}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{\Sigma}_{\mathtt{S},ft}) \,, \quad \text{and} \quad \mathtt{N}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{\Sigma}_{\mathtt{N},ft}) \,, \tag{1.13}$$

where $\mathbf{\Sigma}_{\mathtt{S},ft} \in \mathbb{C}^{M \times M}$ and $\mathbf{\Sigma}_{\mathtt{N},ft} \in \mathbb{C}^{M \times M}$ are the covariance matrices of speech and noise, respectively. The spatial covariance matrix can be factorized into $\mathbf{\Sigma}_{ft} = \sigma_{ft}^2 \boldsymbol{R}_f$, where σ_{ft}^2 are time-varying temporal-spectral variances that describe the temporal-spectral power of sources and $\boldsymbol{R}_f \in \mathbb{C}^{M \times M}$ are time-invariant spatial covariance matrices characterizing the sound propagation process from sources to microphones [27, 35]. The speech posterior can be derived using Bayes' theorem, similar to the single-channel case discussed in Section 1.2.1. Computing the posterior mean of the speech spectral

coefficients also leads to the multi-channel Wiener filter [34]:

$$\mathbf{W}_{ft} = \mathbf{\Sigma}_{\mathtt{X},ft}^{-1} \mathbf{\Sigma}_{\mathtt{S},ft}, \tag{1.14}$$

where the mixture covariance matrix $\mathbf{\Sigma}_{\mathtt{X},ft}$ can be decomposed as $\mathbf{\Sigma}_{\mathtt{X},ft} = \mathbf{\Sigma}_{\mathtt{S},ft} + \mathbf{\Sigma}_{\mathtt{N},ft}$ under the assumption that speech and noise are uncorrelated. The multi-channel Wiener filtering can be achieved through various parameter estimation methods and has been applied in, e.g., audio source separation, blind speech separation, and speech enhancement [36, 37, 27, 10]. For example, these model parameters can be obtained in the maximum likelihood sense using expectation-maximization (EM) update rules [27] or multiplicative update rules [10]. In this thesis, we focus on its application in combination with the VAE and multichannel extensions of NMF [10, 38], as will be discussed in Section 1.4.2. In other words, the multi-channel Wiener filter presented here serves as a basis for extending the VAE-NMF framework considered in Section 1.4 to the multi-channel case [38] [P8]. It is worth noting that this thesis focuses primarily on single-channel speech enhancement, and unless explicitly stated otherwise, all discussions and analyses are conducted in the single-channel context.

Eventually, based on the available prior knowledge, such as the collection of speech and noise data, parameter estimation methods regardless of single-channel and multi-channel applications can be grouped into *supervised learning*, *unsupervised learning*, and *semi-supervised learning*, which will be described in the next section.

1.2.2 Parameter Estimation

By imposing different statistical assumptions on speech and noise, we can develop various clean speech estimators for single-channel and multi-channel application scenarios. It is important to recognize that the effectiveness of statistical speech estimators is closely tied to the accurate estimation of their parameters. In single-channel scenarios, clean speech estimators usually require estimates of temporal-spectral PSDs of speech and noise, while in multi-channel scenarios, it additionally requires accurate estimation of the spatial covariance matrices of speech and noise. Thus, clean speech estimators can be combined with various parameter estimation schemes. For example, the noise PSD can be estimated from the noisy observation by detecting speech activity [7, 39] or more advanced speech presence probability [40, 6]. However, these methods often rely on the assumption that noise is changing more slowly than speech, which is not valid when dealing with highly non-stationary noise interferences. Therefore, statistical model-based parameter estimation relying on simplified assumptions may show limited performance under challenging acoustic conditions. Machine learning-based algorithms can overcome this limitation to some extent by learning discriminative patterns directly from data without making explicit prior assumptions about the acoustic environment. This advantage has led to the integration of representation learning techniques, such as NMF [8, 41, 42, 35, 27, 10], with the statistical speech estimators, as will be detailed in Section 1.4. More recently, DNNs have emerged as a more powerful tool and have revolutionized many application scenarios, including speech signal processing, primarily due to their universal approximation capabilities. DNNs can model complex

non-linear relationships in data more effectively than traditional methods and thus allow capturing intrinsic variations in speech and noise. As a result, deep learning-based methods can improve the overall performance of speech enhancement in acoustically adverse scenarios. The use of DNNs in speech enhancement will be described in more detail in the following sections.

Different types of parameter learning methods can be developed based on specific prior knowledge. Machine learning algorithms typically operate in two steps, i.e., training and testing. Depending on the extent of available training data, these algorithms may be treated differently during the training stage. This flexibility allows the adaptation to various application scenarios. Note that it may also be possible to bypass the training step when no training data is available. In this thesis, we categorize training mechanisms into three groups: *supervised learning*, *semi-supervised learning*, and *unsupervised learning*, based on whether we have access to labeled data. Note that these terms may have different interpretations in different contexts of the literature. For example, obtaining model parameters from isolated source signals is referred to as "unsupervised modeling" in [43]. It is termed "supervised modeling" when additional annotation of isolated sources, such as note information in music signal processing, is provided. In computer vision tasks, it is typically referred to as "semi-supervised learning" when a limited number of labeled samples alongside a large amount of unlabeled data are available [44]. This thesis is concerned with speech enhancement involving speech and noise sources, so here we clarify these terms accordingly to avoid possible confusion. We define these terms based on the nature of the training data available. More specifically, we relate the supervision to whether we can access corresponding speech targets given noisy mixtures. We consider it to be semi-supervised learning when the training stage has limited prior knowledge (e.g., isolated speech or noise) and refer to it as unsupervised learning when no prior knowledge is accessible other than training noisy mixtures. In summary, we distinguish three learning strategies:

- *Supervised learning* relies on parallel noisy-clean speech data, that is, the training process involves using noisy mixtures and their corresponding isolated speech and noise signals. The labeled dataset can be used in a variety of ways depending on the algorithm under consideration. For instance, various training targets can be constructed from a paired synthetic dataset, as will be introduced in Section 1.3.

- *Unsupervised learning* includes methods that rely solely on noisy mixtures during training. It also includes methods that do not require any training data but can derive solutions directly from individual test samples.

- *Semi-supervised learning* lies in between supervised and unsupervised learning. Unlike supervised learning, it cannot access paired noisy-clean speech but can train the model using isolated speech or noise data. This implies that only limited prior knowledge is known, e.g., specific application acoustic scenarios or speaker information, as will be discussed in Section 1.4.

In general, supervised algorithms aim to learn from paired training data and apply it to unseen samples. However, its performance is often related to the quality and diversity of training data and

may exhibit limited generalization ability in acoustic scenarios under-represented by training data. In contrast, unsupervised learning schemes refrain from learning from labeled data but focus on leveraging unlabeled noisy mixtures, which are relatively easier to collect. This may be feasible and preferred when paired training samples from the target domain are difficult or costly to collect. However, gaining high performance in this blind manner remains a challenging research problem. Semi-supervised algorithms acquire knowledge specific to each sound source by analyzing isolated speech or noise data, independent of deterministic mapping relationships established through synthetic noisy-clean speech pairs. In this thesis, we refer to the definitions above when discussing different settings in deep speech enhancement.

1.3 Predictive Speech Enhancement

Deep learning-based speech enhancement algorithms have been an active research topic and currently offer state-of-the-art performance thanks to the powerful non-linear modeling capabilities of DNNs. Among them, deep predictive methods are widely used, aiming to learn the mapping between noisy mixtures and the corresponding clean speech. These algorithms are often trained on a synthetic dataset consisting of numerous paired noisy-clean speech samples covering a wide range of acoustic conditions, such as various noise types and different input SNRs. The supervised predictive scheme may perform well when the prior knowledge acquired from training data matches the test condition. However, an inevitable data mismatch between training and testing may result in degraded speech enhancement performance. The performance gap becomes more prominent when this distribution shift increases. Building upon deep predictive learning, this thesis delves into a new yet crucial aspect: how *uncertainty* is modeled and quantified within a DNN-based speech enhancement framework. This section presents an overview of deep predictive speech enhancement in Section 1.3.1 and then discusses uncertainty modeling in Section 1.3.2.

Deep speech enhancement is often formulated as a regression task to restore clean speech. Alternatively, when dealing with noisy signals intended for the recognition modules of intelligent machines, this problem can optionally be framed as a classification task, in which DNNs can predict the phonemes of a speech utterance, i.e., phoneme recognition [P5]. Therefore, in addition to uncertainty modeling in regression tasks, we also briefly introduce uncertainty modeling in classification tasks in Section 1.3.2.

1.3.1 Deep Predictive Speech Enhancement

DNNs can learn complex patterns from data and can be trained to approximate various training targets [4], allowing them to be flexibly integrated with traditional speech enhancement methods. Moreover, existing work has explored end-to-end speech enhancement systems, where neural networks are used to output speech predictions directly through spectral mapping or indirectly through filter masking. Depending on their application, varying degrees of success have been reported.

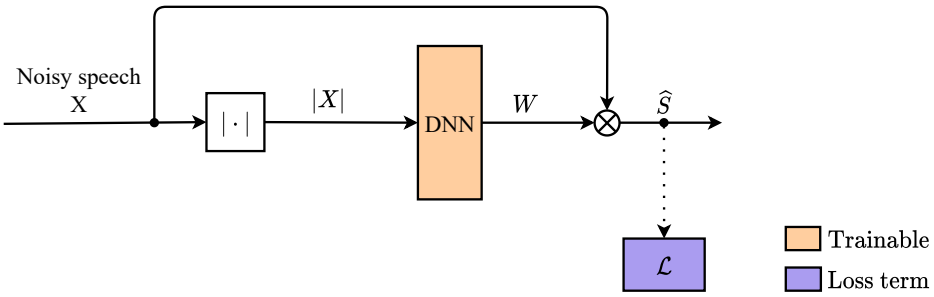


Fig. 1.1 Deep predictive masking-based speech enhancement. The neural network takes as input the magnitude spectrogram, $|X|$, and outputs a filter mask, W, to extract clean speech from the complex spectrogram of the noisy mixture, $\widehat{S}$.

DNNs have been utilized to replace some of the building blocks of traditional speech enhancement methods. To overcome the tracking delay caused by statistical model-based noise PSD trackers [40, 6], Zhang et al. have proposed a DNN-based noise PSD tracker, which utilizes a neural network to estimate the *a prior* SNR required in an MMSE estimate of the noise periodogram [45]. A periodogram is an estimate of a signal's PSD, which describes the power distribution over frequencies. The follow-up work has also proposed to improve it with different training targets and more advanced architectures [46–48]. Furthermore, a neural network-based speech presence probability estimator has been proposed in [49] for robust detection performance and combined with a single-channel multi-frame clean speech estimator, i.e., the multi-frame minimum variance distortionless response (MVDR) filter [50]. The single-channel multi-frame filter extends the well-known multi-channel MVDR filter to single-microphone applications, that is, instead of employing spatial information between multiple microphones, the extended single-channel filter exploits inter-frame correlation to perform noise reduction under the speech distortion-free constraint. In the follow-up work [51], Tammen et al. integrate this multi-frame estimator into a deep learning framework by using DNNs to estimate necessary parameters that are typically difficult to estimate blindly from noisy mixtures [52, 53], such as a prior SNR and highly-varying inter-frame correlation vectors. In contrast to methods that construct intermediate training targets, they optimize the parameters of DNNs in an end-to-end fashion, which is achieved by applying differentiable operations to the intermediate network outputs. Similar ideas can also be found in multi-channel multi-frame approaches [54, 55]. Note that there are no explicit assumptions applied to the network architecture, allowing for the flexible design.

Another research line that dominates predictive learning is masking-based approaches, which aim to learn a mapping between noisy mixtures and filter masks to extract clean speech [56, 4], as illustrated in Figure 1.1. These approaches focus less on statistical assumptions but mainly leverage the data-driven nature of DNNs. By training on a dataset that covers a wide range of acoustic scenarios, DNNs can discover underlying patterns of target sources. Various training targets have been constructed in the time-frequency domain using paired speech and noise signals. For example, the sparsity of the time-frequency representation leads to W-disjoint orthogonality [57], which indicates

that sound sources exhibit less overlap in the spectral domain. This motivates the development of an *ideal binary mask*, which assigns a time-frequency bin to one of the sound sources in the mixture based on the SNR. Besides speech enhancement, this has also been used in DNN-based speech separation, such as in deep clustering [58]. An improved filter mask is an *ideal ratio mask*, which instead of determining speech and noise through a binary decision, assigns a soft value between zero and one to indicate how likely it belongs to a certain source. Other extension includes phase-sensitive mask [59] and complex ratio mask [60], which implicitly take into account phase information. While it seems that the speech enhancement methods considering phase information can naturally outperform magnitude-based counterparts, an empirical study on DNN-based phase-aware speech enhancement presented in [61] demonstrates that the performance is closely related to the chosen frame length. It reveals that the phase-aware algorithm outperforms the magnitude-only counterpart when operating on short-frame time-frequency representations (e.g., a frame length of 4 ms at a sampling frequency of 16 kHz), while the performance is comparable when using long frame lengths, e.g., 32 ms. A similar discussion can also be found in [62]. While phase modeling is an intriguing research topic and may potentially provide additional performance gains when appropriately used [63], it is not the main focus of this thesis, thus refraining from further discussion.

DNN-based speech enhancement may benefit from integrating various engineered features derived from time-frequency representations of noisy mixtures, such as log-compressed spectrograms and pitch-based features [4]. Moreover, inspired by human auditory perception, input features can be computed based on perceptual scales, such as Mel scale, Bark scale, and equivalent rectangular bandwidth scale [4, 64–66]. However, modern network architectures are often designed with millions of parameters, which tend to be over-parameterized. Given the capability to learn complex features, it remains unclear whether heavily engineered features can consistently outperform raw spectrograms. Another crucial aspect of training a network model is the choice of loss function. An appropriate loss function can provide informative gradients during optimization and guide the network model to reach stable local minima in the loss landscape. A diverse range of loss functions has been proposed in speech enhancement [17], including magnitude MSE/mean absolute error (MAE) [56], complex spectral MSE/MAE [67], and their time-domain counterparts [17, 68]. Similar measures can also be computed on the logarithmic scale [17]. Commonly used energy-based losses include signal-to-distortion ratio (SDR) [69] and scale-invariant signal-to-distortion ratio (SI-SDR) [70]. In general, it is not straightforward to assert that one particular cost function is superior to others, as their effectiveness often depends on various aspects involved in the optimization process and the specific network architecture of choice. Moreover, the loss function does not necessarily need to be applied directly to the output of the DNN. As demonstrated in [51], the intermediate estimates generated by the DNN can be used to compute the final speech estimate, to which the loss function is subsequently applied. The essential criterion is that the operations applied to the DNN's output are differentiable, or approximately differentiable such that gradients can be back-propagated.

In addition to input features and loss functions, an important research focus is the design of network architectures that capture the underlying characteristics of clean speech. This is essential

for maintaining the algorithm's robustness in complex and unseen acoustic scenarios. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been empirically shown to be very effective in feature extraction and modeling, and thus have served as basis components for most existing work in the field. CNNs are typically used to extract representative features from data [13], while RNNs, such as its variants long short-term memory (LSTM) and gated recurrent unit (GRU), are often used to model temporal characteristics of signals [71, 12]. Recent work has also attempted to model long-range temporal information utilizing temporal convolutional networks, which introduce the dilation operation to increase the receptive field of CNN layers [45, 51, 72]. Among existing work, one of the most commonly used architectures in deep learning is *U-Net*, which was originally proposed for medical image segmentation [14] and was later successfully applied to various domains [73–76], including speech enhancement [13, 15, 77, 78, 66]. The U-Net architecture typically features an encoder-decoder structure with skip connections linking the two. When handling data of sequential nature, such as speech, RNNs layers can be inserted between the encoder and the decoder to model temporal information [77, 15]. Further extensions include combining U-Net with temporal convolution networks [79], dual-path RNN [80], attention modules [16], and state space models [81]. While recent work has explored improving U-Net with various emerging architectures, it remains unclear which combination of these modules can provide consistently superior performance in a variety of acoustic environments.

Deep learning algorithms offer great flexibility during algorithm development. For example, some methods employ progressive learning [82] or multi-stage learning [83] to perform speech enhancement through multiple steps. This can be interpreted as decomposing a complex regression task into multiple subtasks, where small gains achieved at each step can be accumulated to better the entire task. Furthermore, perceptual instrumental metrics can be used as loss functions to train neural networks. This entails developing differentiable perceptual instrumental metrics, such as speech quality metric-based methods [84] and intelligibility metric-based methods [85]. However, the existing work has also noted the disadvantages of metric-based loss functions, such as producing unnatural speech [86, 87]. The research also focuses on optimization schemes. For instance, several studies have explored the adversarial training scheme provided by generative adversarial networks [88] to improve clean speech estimation [18, 19]. Moreover, learnable transforms have been a major focus in speech enhancement [89] since the exploratory study on speech separation by Luo et al. [72]. To advocate a learnable transform, one may argue that the transform learned from training data is task-specific and likely leads to more discriminative representations. However, it remains an unresolved question whether learned signal encoding outperforms deterministic ones for speech enhancement and separation tasks, and vice versa [65, 90, 91]. Additionally, research has been conducted to develop complex operation-based neural networks [92, 93], but the advantages against real-valued counterparts are still unclear [94].

Next, we will discuss the incorporation of uncertainty modeling into DNN-based predictive speech enhancement. Specifically, We will build on the prevalent predictive masking approaches in the time-frequency domain.

1.3.2 Modeling Uncertainty in Predictive Learning

The previous section discussed key aspects of progress in deep predictive speech enhancement, from input features to optimization techniques. The literature has reported varying levels of success for these factors, depending on the specific application. Deep predictive models are typically used to output a single-point clean speech estimate for each noisy input, but it is important to note that the trained model's predictions may not always be accurate. Indeed, supervised predictive approaches have shown a high dependence on training data [22, 86, 21] and a slight distribution shift in the data during testing may easily cause the model to perform poorly. Therefore, these approaches formulated as a problem with a single output may result in fundamentally erroneous estimates for complex and unseen samples without indicating that the erroneous estimate is uncertain. Orthogonal to the progress outlined in Section 1.3.1, in this thesis, we investigate whether DNN-based speech enhancement algorithms can produce clean speech estimates while at the same time outputting confidence in the predictions, i.e., *uncertainty*, especially for underrepresented noisy samples.

Uncertainty prediction is of great interest for machine learning models since it can inform about the reliability of estimates in the absence of the ground truth, which is usually unavailable to us during inference. Low uncertainty generally indicates a small deviation from the ground truth, while high uncertainty indicates a large estimation error. Thus, a measure that can reflect how large the error margin is can be used to quantify uncertainty; in other words, uncertainty describes how wide the distribution of predictions is [95]. Depending on the task, various uncertainty measures have been employed, such as confidence intervals, entropy, and variance. For example, Depeweg et al. [96] propose to measure uncertainty based on the entropy of the predictive distribution, which represents the information level of random variables. Pearce et al. [97] use confidence intervals in a distribution-free setting, illustrating how certain the estimate is within a certain range. Another widely used measure is variance [73, 74, 98], which is a measure of the spread of a distribution. Some other metrics, which may be task-dependent, are also employed, such as in [99, 100]. Here, we address uncertainty modeling in a probabilistic manner following [74, 101] and use *variance* in regression settings and *entropy* in classification settings.

Sources of Uncertainty

Methods for modeling and predicting uncertainty in machine learning can be quite diverse and often depend on the chosen setting and methodology; here, we focus on uncertainty modeling in the context of DNNs. Before we proceed with methods to capture uncertainty, we clarify the sources of uncertainty. For example, when dealing with neural network-based speech enhancement, one can identify several sources of uncertainty, such as from statistical model assumptions, data collection process (e.g., data quality), network model design, and randomness introduced during optimization. Despite many sources, they are typically categorized into *aleatoric uncertainty* and *epistemic uncertainty* in machine learning when it comes to uncertainty modeling [102, 103, 11].

Uncertainty estimation in deep learning-based methods has gained increasing interest, especially in computer vision tasks [74, 75]. In [74], the authors define aleatoric uncertainty as inherent uncertainty in data, thus also referred to as *data uncertainty*. In image processing tasks, data uncertainty can often be attributed to, e.g., noisy labels or object occlusion. Images of objects without distinct features, or those with reflective surfaces, also pose challenges to the model and can lead to high data uncertainty. This type of uncertainty is characterized by the stochastic dependency between input and output, which can be represented by means of a conditional probability distribution. It is also considered independent of the network model. Additionally, due to its intrinsic uncertainty nature, this type of uncertainty is viewed as irreducible, as discussed in [74]. In contrast, epistemic uncertainty is described as uncertainty arising from the network model weights in [74], also referred to as *model uncertainty*. This uncertainty indicates a lack of knowledge about the network model and can be reduced by training the network model on more data. For example, high epistemic uncertainty may be observed when an under-represented sample is presented during inference. This type of uncertainty can be captured by modeling the parameters of the neural network stochastically in the form of probability distributions, instead of deterministically by point numbers.

However, clearly defining and distinguishing different types of uncertainty is non-trivial, and the boundary between the two may be blurry and can be context- and task-dependent [103, 95]. Thus, simply transferring the definitions outlined above to the speech enhancement setting may be inappropriate. For example, one may analogize DNN-based speech enhancement to an image processing task, where the spectrogram can be treated as an image and each time-frequency bin corresponds to a pixel. The goal is to extract clean speech from the noisy input. Similar to object occlusion in image processing, speech and noise are likely heavily overlapping, which can naturally lead to high aleatoric uncertainty. One solution to improve the performance of separating clean speech from overlapping background noise is to include more training data covering a wide range of noise distortions. For example, for vowel sounds in a magnitude spectrogram, a network model can learn to better recognize harmonic structures with more training data included. This stands in contrast to the non-reducible definition of aleatoric uncertainty provided in [74] since the network model in this task can gain knowledge through more data and then compensate for aleatoric uncertainty [95]. Thus, these observations blur the boundary between the two uncertainties. A similar view has also been presented in [103], where the authors argue that by altering the setting of the problem, such as performing a classification task with an additional feature dimension (e.g., one-dimensional classification vs two-dimensional classification), aleatoric uncertainty can be possibly reduced in a higher dimensional space. Meanwhile, a more complicated model may be required to fit the given dataset, which can potentially increase epistemic uncertainty. This suggests that changing the problem setting may switch one type of uncertainty to the other, emphasizing the difficulty in attempting to have absolute definitions.

Another factor that can increase aleatoric uncertainty is the ambiguity present in the ground truth of the training data, such as the difficulty in labeling occluded object boundaries in image segmentation tasks. In speech enhancement, noisy labels in training data can be related to disturbing

effects in the ground truth clean speech, such as the breathing sounds of speakers and microphone noise. However, this can be largely ignored when using a synthesized dataset containing high-quality clean speech materials. Our analysis often shows that uncertainty is mainly caused by external interfering distortions that we want to remove. Furthermore, we can observe that a single network that models only aleatoric uncertainty can raise high uncertainty for noisy samples not sufficiently represented by training data, such as unseen noise types. This indicates that a single model can also capture the effects of epistemic uncertainty and inform data outliers without applying statistical modeling to the network weights as in [74]. These observations and potential ambiguities again illustrate that distinguishing one uncertainty from another is not straightforward and may be closely related to the task.

In order to avoid potential confusion, in [104, 95], the authors depart from the conventional aleatoric and epistemic terminology and introduce new terms to clarify uncertainty definitions. Here, we want to minimize confusion while using the same terminology, with some necessary clarification. We maintain consistency with most of the existing literature by adhering to the terms aleatoric uncertainty (data uncertainty) and epistemic uncertainty (model uncertainty). However, in our specific task, they may extend beyond the definitions of the inherent randomness in data and the uncertainty of the model weights, respectively. Deep learning algorithms involve many steps that can introduce uncertainty, from data preparation to target prediction. In addition to modeling the inherent randomness of data and uncertainty of the network model weights, uncertainties arising from the training process also contribute to the uncertainty in predictions: If the parameters of a neural network are trained, e.g., using different training data, different initialization schemes and stopping criteria, or a different number of epochs, different parameters result. This is closely related to the convergence of a network model, thus affecting the estimation of uncertainty about the model weights; however, from a practical point of view, this may also have an impact on the estimation of uncertainty inherent in data, as neural networks landed at different local minima of the loss landscape provide different data uncertainty estimates. Therefore, both data and model uncertainty can, to some extent, encompass the uncertainty introduced during the training/convergence process.

When modeling the uncertainty of the network weights, an assumption has been made implicitly: The network model designed is a powerful universal approximator, while the discrepancy caused by the model design is simply ignored. In other words, modeling the posterior of the network weights has been considered sufficiently general to describe not only all model parameter sets but also all possible network models. The uncertainty arising from the model design and selection is practically difficult to model [103]. Here, we share the same view as in [95] that when modeling input-output dependency stochastically in the form of a conditional distribution and leveraging a DNN to perform the estimation, the variance associated with the prediction captures not only the inherent randomness in data but also limitations of the model. Consequently, a single model (without modeling the network weights probabilistically) has the potential to elevate uncertainty estimates in situations where the model lacks knowledge about specific test samples.

In a nutshell, we employ the term *aleatoric uncertainty* or *data uncertainty* to refer to the uncertainty that stems from the stochastic dependency between input and output as well as the limitations of manually designed network architectures. We use the term *epistemic uncertainty* or *model uncertainty* to refer to the uncertainty of the model parameters. Importantly, uncertainty arising from the optimization process contributes to both sources of uncertainty.

Estimating Data Uncertainty

Having discussed the sources of uncertainty in deep learning-based speech enhancement, we first discuss how to estimate the uncertainty in the predictions arising from data uncertainty. Given the stochastic dependency between input and output represented by a conditional distribution, data uncertainty can be quantified by estimating this full predictive distribution. Specifically, with the speech-plus-noise assumption as in section 1.2, one can derive a closed form of the speech posterior $p(S|X)$. For example, a complex Gaussian posterior can be derived based on the common complex Gaussian priors of speech and noise, see equation 1.4. By estimating the statistical moments of this distribution, the predicted mean serves as an estimate of the speech target, while the associated variance can be used to quantify data uncertainty. Essentially, the predicted mean is the MMSE estimate of clean speech, which in this case, is equivalent to the MAP estimate since the posterior is unimodal and symmetric.

Although in principle it is possible to use tools, such as noise and speech PSD trackers, to estimate the full posterior distribution, here we make use of DNNs due to their powerful non-linear modeling capability. A standard neural network that outputs a point estimate for each input noisy sample does not allow the estimation of predictive uncertainty. For instance, a DNN trained by minimizing the widely-used loss function MSE can be interpreted as assuming homoscedastic uncertainty, which implies a constant variance that is not explicitly estimated. To capture input-dependent data uncertainty, we instead predict the full posterior distribution, and in the case of the Gaussian distribution, a simple solution is to split the output layer into two layers to estimate the mean and its associated variance, following $p(S|X) \sim \mathcal{N}_c(\mu(X), \lambda(X))$, where $\mu(\cdot)$ and $\lambda(\cdot)$ denote the mean and variance functions parameterized by θ. Given a training dataset that contains paired noisy-clean speech complex spectrograms $\mathcal{D} = \{(S_{11}, X_{11}), ..., (S_{FT}, X_{FT})\}$, the network weights θ can be obtained by maximum-likelihood estimation, with the loss function:

$$\widetilde{\theta} = \arg\min_{\theta} \mathcal{L}(\theta) = \arg\min_{\theta} -\frac{1}{FT} \sum_{f,t} \log(p(S_{ft}|X_{ft}, \theta)). \tag{1.15}$$

In the speech enhancement task, DNNs output an estimate for each input time-frequency bin. The index ft is thus omitted when no confusion occurs. While the loss function is typically computed by averaging over time-frequency bins of the input spectrogram, a well-designed DNN can effectively exploit both temporal and spectral correlations in the input.

Eventually, in situations where the network model has difficulty in predicting certain samples, either due to inherent stochastic dependency or due to limitations in the network architecture and its training procedure, the network model may exhibit high uncertainty as an indicator that the predictions may be inaccurate. However, training a probabilistic model is not a trivial task. For example, the ground truth of uncertainty is not readily available, making uncertainty estimation an unsupervised task with an unspecified search space. Also, in order to appropriately capture a full distribution parameterized by the mean and variance, the neural network may require to be trained on a sufficient amount of data. Furthermore, once we have reliable estimates of data uncertainty, the question that may arise is how we can leverage them to further enhance the estimation of clean speech. These questions are explored and addressed in our contribution [P2], where we employ an approximate MAP estimator of spectral magnitudes that explicitly requires variance estimates to further refine speech estimates [105]. The proposed joint scheme establishes an interesting connection between the complex spectral domain and the magnitude domain based on complex Gaussian assumptions. The resulting joint loss objective has also been shown to be effective in stabilizing the training process.

Estimating Model Uncertainty

In the previous section, we discussed how to employ a DNN to capture data uncertainty. However, directly estimating the full conditional distribution $p(S|X)$ based on a single deterministic neural network is agnostic to model uncertainty. To capture uncertainty in predictions due to model uncertainty, one can utilize Bayesian deep learning, which provides a principled tool to model uncertainty in neural networks. Rather than assuming deterministic neural network parameters, Bayesian deep learning treats the network parameters θ as random variables and models the network weights stochastically by placing a distribution over them. Thus, the task of estimating model uncertainty boils down to performing Bayesian inference, i.e., deriving the posterior over the network weights $p(\theta|\mathcal{D})$ that captures model uncertainty. Formally, with a proper prior $p(\theta)$ and likelihood $p(\mathcal{D}|\theta)$, one may derive a posterior distribution using Bayes' rule:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}, \tag{1.16}$$

where $\mathcal{D}$ denotes a dataset comprising paired noisy-clean speech samples. However, for neural networks with millions of parameters, it is computationally intractable to compute an analytic form of the posterior distribution (e.g., due to difficulty in computing the integration in the evidence $p(\mathcal{D})$). Hence, various approximation methods have been proposed.

To perform Bayesian posterior inference, early work has been focused on sampling-based methods, e.g., Markov chain Monte Carlo (MCMC) methods, which construct a Markov Chain with the posterior distribution as its stationary distribution. In order to scale to large datasets, stochastic gradient versions have been investigated [106–108], where a subset of data, i.e., a minibatch of data, is employed for gradient computation instead of the whole dataset. However, MCMC methods are computationally

inefficient and face challenges in handling high-dimensional functional spaces, especially for DNNs with a large number of parameters [75, 95].

Another method of approximate inference is variational inference, where one can approximate the complex posterior distribution $p(\theta|\mathcal{D})$ with a simpler variational distribution $q(\theta)$, such as Gaussians [109, 110]. The discrepancy between these two distributions can be measured through a predefined metric, which turns the approximation problem into an optimization problem. Among them, a predominant measure is the Kullback-Leibler (KL) divergence. Specifically, with reverse KL divergence $\mathrm{KL}[q(\theta||p(\theta|\mathcal{D}))]$, the problem can be formulated as maximizing the evidence lower bound (ELBO). Then, intractability due to the absence of the analytical solution of the posterior distribution is avoided, and the optimization with respect to the ELBO is mainly to find the approximate variational distribution $q(\theta)$. A detailed discussion of this approximation will be given mathematically in Section 1.4.1. The main concern is that the terms in the ELBO require computing the expectation with respect to the variational distribution. This may also not be computed analytically but can be approximated via Monte Carlo sampling [99]. To embed this into the context of neural networks optimized using stochastic gradient optimization, another problem to consider is how to make the sampling process differentiable. To solve this, Kingma et al. [109] proposed a reparameterization technique, which is a differential transformation. This transformation allows the backpropagation of the gradient through the parameters of interest while at the same time injecting stochasticity through a random sampling process. While Kingma et al. apply this technique to autoencoders and randomize their hidden variables [109], the follow-up work [110] extends this to capture the probability distribution of neural network weights in a much higher dimensional space and diversifies the variational distribution with a Gaussian mixture model. While variational inference tends to be more computationally efficient than sampling-based methods [111, 112], there are some challenges highlighted in the literature [112, 101]. For example, it is difficult to determine a proper common prior, considering various possibilities of network design; optimizing the objective function KL divergence may not be the optimal choice and alternative divergence measures may improve overall approximation quality [112, 99]; independence assumptions have typically been made among the network weights to facilitate derivation of solutions for large DNNs, but this may potentially limit the approximation performance. Furthermore, depending on the chosen variational distribution, Bayesian neural networks may greatly increase the number of parameters. For instance, with a Gaussian distribution fully described by a mean and variance, the number of parameters is doubled compared to the corresponding deterministic countpart. In practice, it may not be necessary to introduce stochasticity to all parameters; instead, one can selectively randomize specific parts of a DNN [73, 113].

Despite potential limitations, some methods performing variational inference have shown promising uncertainty estimation performance for large-scale DNNs [75, 74]. For example, Gal et al. [114] approximate the posterior distribution with a Gaussian mixture model (i.e., a mixture of two Gaussians) and build an interesting connection between variational inference and the dropout regularization technique [115]. This method is referred to as *Monte Carlo (MC) dropout*. Sampling from the

posterior distribution of the network weights at inference time is achieved by activating dropout. A single forward pass generates a set of network parameters. By performing multiple forward passes, we generate multiple sets of parameters, simulating sampling from a network parameter posterior distribution. Given the widespread use of dropout in architectural design, MC dropout can scale easily to large datasets and network architectures. Recent studies have also attempted to adopt MC dropout in speech processing tasks, such as speech recognition [116–118] and speech emotion recognition [119].

While Bayesian inference provides an elegant tool to reason about the posterior distribution of the network weights, model uncertainty can also be captured by an ensemble of deterministic networks, i.e., ensembling [98]. The motivation is that a network with millions of parameters gives rise to a multi-mode function space and a network trained multiple times may land in different local minima of this high-dimensional space due to randomness introduced during the optimization process (e.g., initialization, data shuffling). An ensemble of network models can then be viewed as samples obtained from some network parameter posterior distribution and is thus also considered equivalent to approximate Bayesian inference [75, 120]. Consequently, while Bayesian deep learning techniques allow a trained network to be sampled from a local mode in the high-dimensional space, an ensemble of networks trained with randomness may allow sampling from different local minima, as empirically shown in [121]. This hypothesis has also been supported by comprehensive experimental comparison reported in [75], in which Deep ensembles [98] outperform other Bayesian inference methods, leading to state-of-the-art performance in uncertainty estimation. Various ensembling techniques have been designed in the literature [122, 123]. For example, an ensemble of network models can be obtained through different random initialization seeds, i.e., training the same network model multiple times on the same amount of data using random initialization, such as Deep ensembles [98]. Furthermore, one can create network variation by bagging (also called bootstrapping [123]), which uses the same initialization seed but trains the network model on different subsets of the training data through resampling. A comparison presented in [122] has shown that for large models trained with a large amount of training data, random initialization is effective and preferred over bagging in terms of model diversity. Specifically, it has been pointed out in [122, 98] that training on a resampled version of the dataset leads to worse performance, possibly due to that data resampling reduces the number of unique samples. This observation has also been empirically evaluated in [123].

Overall, the true posterior distribution of network weights can be approximated through **1)** sampling-based methods, **2)** approximate variational inference, and **3)** ensemble-based methods. With M network realizations obtained from an approximate posterior network parameter distribution, we can generate M output predictions for each input sample. Uncertainty in predictions due to model uncertainty can be empirically quantified by the variance in these output predictions

$$\begin{aligned} \widetilde{S} &= \frac{1}{M}\sum_{m=1}^{M}\widetilde{S}_{\theta_m}, \\ \widetilde{\Sigma} &= \frac{1}{M}\sum_{m=1}^{M}|\widetilde{S}_{\theta_m}-\widetilde{S}|^2, \end{aligned} \tag{1.17}$$

where $\widetilde{S}_{\theta_m}$ denotes a clean speech estimate using the network model with parameters θ_m. $\widetilde{S}$ and $\widetilde{\Sigma}$ represent the average clean speech estimate and the associated model uncertainty, respectively. Note that there is a clean speech estimate and an associated variance estimate for each ft-th bin, and the index ft is omitted for brevity.

Previous studies on model uncertainty have been focused on tasks other than speech enhancement, e.g., [73, 104, 119, 124, 116–118]. Yet, questions such as how reliable and accurate the estimates of epistemic uncertainty are in speech enhancement, and how modeling epistemic uncertainty affects enhancement performance, have not been addressed. Hence, we study these questions in [P2] and investigate two Bayesian deep learning techniques: MC dropout and Deep ensembles, to capture model uncertainty in clean speech estimation because of their efficiency in approximating Bayesian inference and scalability to large neural networks.

Estimating Overall Predictive Uncertainty

Having discussed how to quantify data uncertainty and model uncertainty in the context of deep speech enhancement, in this section, we illustrate how to estimate overall predictive uncertainty, which reflects both sources of uncertainty. To achieve this, we can inject the speech posterior distribution into approximate Bayesian inference methods. In the case of MC dropout, we can use a neural network with dropout regularization to estimate the full speech posterior distribution by minimizing the objective (1.15); similarly, for Deep ensembles, we train M network models, each of which is optimized using the loss function (1.15). At inference time, the predictive distribution is obtained by marginalizing out the posterior network parameter distribution:

$$p(S|X,\mathcal{D}) = \int p(S|X,\theta)p(\theta|\mathcal{D})d\theta. \tag{1.18}$$

Due to computational intractability, we approximate the expectation with a MC estimator:

$$p(S|X,\mathcal{D}) = \frac{1}{M}\sum_{m=1}^{M} p(S|X,\theta_m), \quad \text{with} \quad \theta_m \sim q(\theta), \tag{1.19}$$

where θ_m denotes samples from an approximate posterior distribution $q(\theta)$.

With the law of total variance [95, 99, 101], the total variance $\mathrm{V}(S|X)$ can be decomposed into the variance of speech posterior means and the expectation of speech posterior variances:

$$\underbrace{\mathrm{V}[S|X]}_{\text{Overall Predictive Uncertainty}} = \underbrace{\mathrm{V}_{p(\theta|D)}[\mathrm{E}[S|X,\theta]]}_{\text{Model Uncertainty}} + \underbrace{\mathrm{E}_{p(\theta|D)}[\mathrm{V}[S|X,\theta]]}_{\text{Data Uncertainty}}. \tag{1.20}$$

Therefore, the former approximates uncertainty arising from model uncertainty, while the latter approximates uncertainty in predictions due to data uncertainty. Similarly, since these quantities can not be computed analytically, they are estimated using MC estimators as in (1.19). The approximate

overall predictive variance is given by:

$$\widehat{\Sigma} = \frac{1}{M}\sum_{m=1}^{M}(|\widetilde{S}_{\theta_m} - \widetilde{S}|^2 + \widetilde{\lambda}_{\theta_m}). \tag{1.21}$$

where $\widetilde{\lambda}_{\theta_m}$ denotes the speech posterior variance estimated by the DNN with parameters θ_m. A clean speech estimate can be obtained by averaging the multiple predictions as in (1.17). The index ft is omitted for brevity.

Bayesian inference provides principled methods to model uncertainty. However, one downside of the approximate inference methods outlined above is their computational inefficiency. During inference, the computational effort increases in proportion to the number of realizations M. While ensemble-based methods, such as Deep ensembles, deliver superior results in uncertainty estimation, training M individual network models makes them less efficient in terms of memory and storage. Therefore, an effective alternative solution might be to generate multiple predictions with a single forward pass through the network, thus bypassing the need for an extensive sampling process. This results in our contribution in [P3], in which we combine the powerful nonlinear modeling capabilities provided by neural networks with super-Gaussian speech priors as a way to improve the robustness of the algorithm as well as to capture predictive uncertainty efficiently.

Uncertainty Estimation in Classification Tasks

In previous sections, we have discussed modeling predictive uncertainty in regression settings. We now present how to estimate data uncertainty and model uncertainty in classification tasks, in which DNNs are employed to estimate discrete class labels. More specifically, network models in classification settings are used to estimate the probability of an input sample belonging to a specific class.

Similar to the regression setting, the network predicts the class variable's posterior distribution, which is usually modeled as a categorical distribution, i.e., a discrete random variable with multiple possible outcomes. We can denote a discrete random variable that can take values from I categories by $C \in \{1, \cdots, I\}$. We denote p_i as the probability of classifying the input data as class i. The sum of p_i is constrained to be 1, i.e., $\sum_{i=1}^{I} p_i = 1$. Hence, given the input X, the classification task is to find the mode of the posterior:

$$p(C|X) = \prod_{i=1}^{I} p_i^{\mathbb{I}(C=i)}, \tag{1.22}$$

where $\mathbb{I}(\cdot)$ is an indication function equal to 1 when $C = i$ and 0 otherwise. For deep learning multi-class classification, the DNN can output a logit score for each class, which can then be normalized using a Softmax activation function. Given an independently and identically distributed data set consisting of input data and corresponding class labels, $\mathcal{D} = \{(C_1, X_1), \cdots, (C_N, X_N)\}$, a network model parameterized by θ can be trained by minimizing the negative log-posterior, which is the well-known cross-entropy loss function.

By estimating the parameters of a categorical distribution using a deterministic network, the entropy of a discrete distribution can capture uncertainty in predictions due to data uncertainty [101]. However, this does not take into account model uncertainty. To this end, we can adopt Bayesian deep learning techniques to sample an ensemble of network parameters $\{\theta_m\}_{m=1}^{M}$. Thus, multiple network realizations can generate multiple sets of logits for an input sample. The expectation of the resulting entropy with respect to the network posterior quantifies the data uncertainty, i.e., $E_{p(\theta|\mathcal{D})}\left[\mathcal{H}\left[p(C|X,\theta)\right]\right]$ [99], where $\mathcal{H}[\cdot]$ represents the entropy. In addition, the predictive distribution can be computed by marginalizing out the network parameter posterior as in (1.18), i.e., $E_{p(\theta|\mathcal{D})}\left[p(C|X,\theta)\right]$. In practice, this can be approximated by MC approximation due to computational intractability, i.e., the average of M predictions. Consequently, the entropy of this expected distribution quantifies the overall predictive uncertainty [101, 99], i.e., $\mathcal{H}\left[E_{p(\theta|\mathcal{D})}\left[p(C|X,\theta)\right]\right]$. Eventually, by computing the difference between the total predictive uncertainty and data uncertainty, we can obtain uncertainty in predictions arising from model uncertainty, which quantifies the spread of predictions from an ensemble of models [101]:

$$\underbrace{\mathcal{I}[C,\theta|X,\mathcal{D}]}_{\text{Model Uncertainty}} = \underbrace{\mathcal{H}\left[E_{p(\theta|\mathcal{D})}\left[p(C|X,\theta)\right]\right]}_{\text{Overall Predictive Uncertainty}} - \underbrace{E_{p(\theta|\mathcal{D})}\left[\mathcal{H}\left[p(C|X,\theta)\right]\right]}_{\text{Data Uncertainty}}, \tag{1.23}$$

Where $\mathcal{I}[\cdot]$ denotes the mutual information operator. Model uncertainty is captured through the mutual information between the network parameters θ and the class C, which represents the amount of information that can be gained about one variable given the other variable [103, 125].

Reliable uncertainty estimates can indicate when a model lacks confidence in its predictions. We want to explore further uses of uncertainty estimates to improve speech enhancement tasks in both regression and classification settings. In our work [P4], we seek to improve the noise robustness of speech enhancement algorithms using unlabeled data from the *target domain* to which the model is applied. This is achieved through a process known as *uncertainty-based filtering*, where we filter out low-quality training samples based on uncertainty estimates and adapt the model to high-quality speech samples. Another application we explore is audio-visual phoneme recognition. While complementary features of the visual modality can improve overall performance, unreliable visual input may result in degraded performance that may be even worse than methods based solely on the audio modality [126] [P5]. Existing works in audio-visual speech recognition have proposed to address visual input distortions by jointly processing modality-specific posteriors and reliability measures using a fusion network [127, 128] or by introducing more carefully designed network architectures [126]. Our work in [P5] studies a uncertainty-based fusion scheme for audio-visual phoneme recognition, where a simple uncertainty-weighted combination is performed at the output level without relying on an additional post-fusion DNN. Moreover, the uncertainty-based fusion strategy uses only clean visual data during training, making it video corruption-agnostic, and is thus expected to generalize to various visual distortions. It allows the model to identify when predictions from a particular modality might be less reliable and determine the extent to which the decision depends on each modality.

1.4 Generative Speech Enhancement

The previous section is focused on deep predictive approaches, where we rely on paired noisy-clean speech data and train neural networks to learn a deterministic mapping between the noisy input and clean speech. While supervised predictive methods have achieved impressive results under matched conditions, their ability to generalize to unseen conditions is limited. The distribution mismatch between training and testing data may lead to performance degradation. However, recent work leveraging generative models has shown the potential to narrow the performance gap between matched and mismatched scenarios. Generative models aim to learn, implicitly or explicitly, probability distributions of complex data, thus capturing inherent signal characteristics. In the time-frequency domain, this includes particularly temporal-spectral features. The learned statistical models can adapt to various acoustic conditions and demonstrate high performance in the presence of noise disturbances unseen during training.

The variational autoencoder (VAE) is a powerful deep generative model that has gained widespread attention in machine learning [109]. It is trained to capture complex features in high-dimensional data such as speech and images by learning low-dimensional latent representations. This makes it suitable for many tasks, such as data compression and representation learning. Recent work has explored its application in speech enhancement [21, 129, 20, 38]. For example, the VAE has been used to learn a prior distribution of clean speech. This is then combined with an untrained non-negative matrix factorization (NMF)-based noise model to perform speech enhancement using a Bayesian inference algorithm, such as Monte Carlo expectation maximization (MCEM) [20]. In contrast to deep predictive models, the original VAE-NMF methods in [21, 20] do not require paired noisy-speech data, but only learn prior speech knowledge on isolated clean speech data. This makes the algorithms *semi-supervised*.

In this chapter, we first describe the probabilistic tools to model speech and noise signals. This is followed by the MCEM, based on which we estimate unknown parameters at inference time.

1.4.1 Non-Negative Matrix Factorization and Variational Autoencoder

Non-Negative Matrix Factorization

NMF is a widely used technique that can learn the intrinsic characteristics of given data [8]. NMF aims to factorize a non-negative matrix into the product of two lower-dimensional matrices, under the constraint that these matrices have no negative values. For example, NMF can decompose an image by identifying basic parts of its composition. NMF has also been a prevailing technique to model speech and audio signals. A common example is to take some preprocessing to convert the complex STFT coefficients to non-negative values, e.g., taking the absolute value or the square. NMF can then be used to find representative features that form the input magnitude spectrogram or periodogram.

Formally, given a non-negative matrix $\mathbf{Y} \in \mathbb{R}_+^{F \times T}$, NMF decomposes it into two non-negative matrices $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K \times T}$, such that their product approximates the input $\widehat{\mathbf{Y}} \approx \mathbf{WH}$. Let $\widehat{\mathbf{Y}}$

denote an estimate of the input $\mathbf{Y}$. Here, F and T are the number of rows and columns of the input matrix (e.g., corresponding to the number of frequency bins and time frames of the input spectrogram), indexed by f and t, respectively. $\mathbb{R}_+$ denotes the positive real values. Essentially, the data matrix $\mathbf{Y}$ can be interpreted as a linear combination of the columns of the matrix $\mathbf{W}$, with weights specified by the elements in the matrix $\mathbf{H}$. For illustration, we take one column vector from $\mathbf{Y}$, denoted as $\mathbf{Y}_t \in \mathbb{R}_+^F$, and the corresponding vector in $\mathbf{H}$, represented by $\mathbf{H}_t \in \mathbb{R}_+^F$, visualized in Figure 1.2. The approximation is then given by $\widehat{\mathbf{Y}}_t \approx \mathbf{W}\mathbf{H}_t$, which indeed gives a weighted sum of the columns of $\mathbf{W}$, with weights specified by the entries in $\mathbf{H}_t$. Thus, the matrix $\mathbf{W}$ is referred to as a *dictionary* matrix or *template* matrix, which contains the columns that serve as basis components (or templates). The matrix $\mathbf{H}$ is referred to as an *activation* matrix, which describes the temporal activity of input. K indicates the size of the dictionary and can be set much smaller than F and T (i.e., $K \ll F$ or T). Hence, this gives a compressed version of the input matrix, and NMF can also be seen as a dimension reduction technique [8]. However, in some applications [130, 131], an overcomplete dictionary is desired, where the dictionary size K is larger than F or T. This approach is often coupled with a sparsity constraint to ensure that the dictionary captures underlying features instead of memorizing input details. Furthermore, the non-negative constraint can help prevent dictionary elements from canceling each other out, thereby encouraging the learning of both meaningful and interpretable basis representations of the input [8, 132]. Its applicability has been demonstrated in various domains. This thesis specifically focuses on its application in speech and audio signal processing.

Given the approximation problem defined above, the decomposition can be mathematically solved by minimizing an error measure:

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \mathbb{D}(\mathbf{Y}|\mathbf{W}\mathbf{H}) \quad \text{and} \quad \mathbb{D}(\mathbf{Y}|\mathbf{W}\mathbf{H}) = \sum_{f=1}^{F}\sum_{t=1}^{T} \mathbb{D}([\mathbf{Y}]_{ft}|[\mathbf{W}\mathbf{H}]_{ft}). \tag{1.24}$$

$\mathbb{D}(\cdot|\cdot)$ denotes a cost function that measures the discrepancy of approximation, and $[\cdot]_{ft}$ denotes the ft-th element of the matrix. Common cost functions include the Euclidean distance and the KL divergence, as introduced by Lee et al. in [8]. Since both measures are non-convex in the joint optimization of $\mathbf{W}$ and $\mathbf{H}$, it is not guaranteed to find a global minimum. One can resort to iterative optimization, which optimizes one variable while fixing the other, allowing NMF to converge to a local minimum. However, due to its non-convex nature, the NMF algorithm with different initializations may converge to different local minima [133]. The optimization process can be implemented using simple techniques such as gradient descent to find these local minima [8]. Interestingly, Lee et al. have shown that when carefully choosing the learning rate to be adaptive, the gradient descent-based additive update rule can be transformed into an easy-to-implement multiplicative update rule. The multiplicative rule can also be derived by using auxiliary functions [8]. This further reduces the burden of hyperparameter search during optimization, such as avoiding the need to search for the learning rate. The following works have also extended NMF with various cost functions [134, 135, 42, 41]. For example, Févotte et al. in [41] propose to measure the approximation discrepancy with the

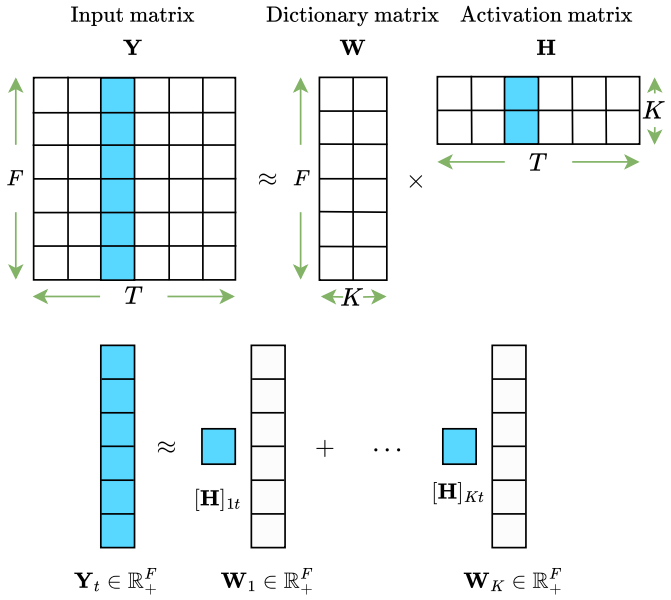


Fig. 1.2 Illustration of NMF that decomposes an input matrix $\mathbf{Y}$ into a dictionary matrix $\mathbf{W}$ and an activation matrix $\mathbf{H}$.

scale-invariant Itakura-Saito (IS) divergence and, similar to the Euclidean distance and KL divergence, provide easily implemented multiplicative update rules. It is interesting in speech and audio signal processing because the IS divergence depends only on the ratio between the true value and the approximate value. Therefore, it treats different frequency components equally and disregards energy distribution over frequency, [41, 10]. This makes the algorithm robust to energy variations and prevents quieter frequencies from being dominated by louder ones. In general, the choice of the cost function may depend on the specific application requirements.

Error measures of NMF can also be defined using appropriate statistical models [136, 10]. For example, optimizing NMF with the Euclidean distance can be interpreted as maximizing the log-likelihood, which has the form of a Gaussian with mean defined as the NMF approximation and a constant variance. Also, the optimization based on the KL divergence is equivalent to performing maximum likelihood estimation using a Poisson distribution [10]. Furthermore, by assuming the complex-valued STFT coefficients complex Gaussian-distributed and approximating the variances with the NMF approximation $\widehat{\mathbf{Y}}$, maximum likelihood estimation with respect to the matrices $\mathbf{W}$ and $\mathbf{H}$ leads to the IS divergence. Note that to strictly establish this equivalence, the decomposition needs to be applied to the periodogram [10]. Furthermore, Févotte et al. in [41] interpret the IS divergence-based optimization as the maximum likelihood estimation in Gamma multiplicative noise. Similarly, the probabilistic interpretation has also been extended to multichannel signals in [10]. In the multichannel case, optimizing IS divergence can be associated with a multivariate complex Gaussian

distribution, where the dimension of the covariance matrix is equal to the number of microphones. Euclidean distance-based NMF can also be derived based on a univariate complex Gaussian. Overall, optimization based on the cost functions may align with certain statistical assumptions, building interesting connections with probabilistic modeling. We will introduce later in Section 1.4.2 that the probabilistic interpretation allows NMF to be combined with DNN-based latent variable models, such as VAE, resulting in statistically interpretable deep generative speech enhancement algorithms [20, 21].

NMF has been widely used to model signals that exhibit structured patterns in source separation [10, 130, 137], speech recognition [138–140], and speech enhancement [141–143]. To effectively separate two or more sources in a mixture, each source (e.g., speech or background noise in speech enhancement) can be modeled by a separate NMF model. These unknown parameters are determined under the constraint that the sum of the approximations of multiple sources is optimized to fit the input data. The target source can then be approximated by computing the matrix product of the corresponding dictionary and activity matrices. Moreover, a post-processing step can be applied to refine the target source estimates. For example, in speech enhancement, instead of approximating through direct matrix factorization, one can obtain clean speech estimates by a Wiener-like filter constructed using the estimated variances of speech and noise [10].

Various training schemes have been derived depending on prior knowledge of the data. For example, when no training data is available [141], the dictionary and activation matrices for speech and noise can be learned online during inference. If all source data is available, each NMF model can be pre-trained on the isolated data. During testing, the pre-learned dictionary matrices are fixed, and only the corresponding activation matrices are optimized to account for the temporal activity of individual signal sources in the input [130]. When only a subset of the source data is available beforehand, the corresponding NMF models can be pre-trained, while the remaining source models are adapted to fit the test data. This allows for flexible integration of incomplete prior knowledge of the data into the model. For example, in speech enhancement, a dictionary matrix of an NMF-based speech model can be pre-learned to capture phonetic information [144, 145]. At test time, its activation matrix and all the unknown noise model parameters are learned on the fly from the input. This training scheme is noise-agnostic and thus expected to generalize across various application scenarios. In contrast, if we have prior knowledge of application scenarios, we can anticipate certain types of noise and thus train a noise dictionary matrix accordingly [1, 146]. For example, this idea has led to various ego-noise reduction algorithms [1]. Ego-noise can be effectively modeled by dictionary-based methods due to its distinct temporal-spectral harmonic structures [9, 147, 148]. The dictionary learning can be further regularized by incorporating information from other modalities, such as motor data [149, 150].

We discussed that the two constituent matrices in NMF have unique functional roles: the dictionary matrix $\mathbf{W}$ captures underlying patterns in the data while the activity matrix $\mathbf{H}$ describes the temporal activity. This interpretability can be leveraged to design efficient real-time systems [130, 143, 141], which are of interest in various speech-based interactive scenarios. It can be observed that each input feature vector is treated independently in the original NMF formulation, as illustrated by $\widehat{\mathbf{Y}}_t \approx \mathbf{W}\mathbf{H}_t$ in

Figure 1.2. By learning the corresponding dictionary matrices beforehand and fixing them at test time, only a single column of the activation matrix $\mathbf{H}_t$ needs to be optimized for each input feature vector (e.g., a magnitude/power spectrum vector). This allows us to perform efficient frame-based speech enhancement and source separation [130, 143]. The dictionary matrix $\mathbf{W}$ must be learned online when no prior data are available. This causes difficulty in performing speech enhancement or separation on a frame-by-frame basis. However, it is possible to reduce latency by learning a dictionary matrix of unknown sound sources over a certain number of frames (e.g., with a sliding window), instead of an entire utterance. Nevertheless, dictionary learning requires analyzing a large number of frames to capture underlying features. Thus, a trade-off must be made between dictionary learning and the speed of inference. Eventually, the overall latency is related to the length of the sliding window plus the processing time for the matrix optimization.

To successfully separate different signals in a mixture, it is essential to have a high-quality representative dictionary matrix for each sound source, such that a linear combination of the templates in the dictionary can accurately describe the corresponding target signal. One crucial factor in learning representative basis components is the appropriate selection of dictionary size. It is often set to strike a balance between being oversized or undersized [138]: an undersized dictionary may have difficulty in learning underlying features and representing the complexity of high-dimensional signals, which potentially affects the overall quality of signal reconstruction, while an over-complete dictionary may overfit the training data, which potentially degrades its generalization performance. However, in a setup where the dictionaries are learned in advance during training and are expected to generalize to new unseen samples, a large dictionary size may be required [151, 130]. For example, the dictionaries with sufficiently large sizes can capture diverse variations in speech and noise signals, enabling generalization to unseen speakers and a wide range of noise conditions. This is typically combined with a sparsity constraint imposed on the activity coefficients, such that only a subset of templates are employed to reconstruct source signals [142, 130]. The sparsity constraints can alleviate overfitting and make the learned features more discriminative. In contrast, exemplar-based algorithms solve the problem differently. They do not require a typical training stage to obtain a representative dictionary matrix but construct one by sampling observations from the training data [131, 140]. Despite imposing a sparsity constraint, these methods designed with a large number of templates can pose challenges in terms of storage requirements. Additionally, the dictionary matrices for different sources need to be sufficiently independent, e.g., speech templates should not capture noise features. If the target signal can be approximated using basis templates from other source signals, unwanted interference may be leaked into the reconstructed target source, resulting in performance degradation. Research has been conducted to distinguish distinctive temporal structures between source signals and incorporate temporal dependency across time frames to improve signal modeling quality and thus speech-denoising performance [152, 145].

Recent work has explored combining representation learning capabilities of NMF with non-linear modeling capabilities of DNNs. Particularly, it has been proposed to replace the NMF-based speech model with a VAE, showing superior performance than a fully DNN-based predictive approach [20]

and a fully NMF-based baseline [21]. We will introduce the VAE model in the next subsection and present how the VAE-NMF model can be applied to speech enhancement in Section 1.4.2.

Variational Autoencoder

For a given dataset with observations $\mathbf{y}$, the VAE aims to model its underlying distribution $p(\mathbf{y})$, which is often complex when it comes to high-dimensional data such as speech and images. While the true data distribution is typically unknown, it can be approximated by $p_\theta(\mathbf{y})$, with distribution parameters θ. The optimization goal at the learning stage is to find the parameters θ such that $p_\theta(\mathbf{y})$ can describe the data sufficiently:

$$\widetilde{\theta} = \arg\max_{\theta} \log p_\theta(\mathbf{y}). \tag{1.25}$$

It is assumed that there exist associated hidden random variables $\mathbf{z}$ and that the data is generated conditional on the latent variables $\mathbf{z}$, i.e., $p_\theta(\mathbf{y}|\mathbf{z})$. The marginal distribution is then written as:

$$p_\theta(\mathbf{y}) = \int p_\theta(\mathbf{y},\mathbf{z})d\mathbf{z} = \int p_\theta(\mathbf{y}|\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z}, \tag{1.26}$$

where in Bayesian inference, $p_\theta(\mathbf{y})$ is called the evidence or the marginal likelihood and $p_\theta(\mathbf{z})$ is the prior. The latent variable model provides an effective framework for modeling complex data, and even if we choose simple prior and conditional distributions, the marginal distribution can be flexible and very complicated, as discussed in [153]. A simple example is the Gaussian mixture model [33], which includes the discrete hidden variable $\mathbf{z}$ to indicate the Gaussian component from which the data is generated. Moreover, when the latent variable $\mathbf{z}$ is continuous [153], it can be seen as a mixture model comprised of infinitely many conditional distributions. Therefore, the expressive latent variable model is powerful and potentially capable of approximating high-dimensional complex data distributions.

Note that the marginal distribution in equation (1.26) is computationally intractable, that is, the integral makes it challenging to find a closed-form solution [33]. In practice, it is often more feasible to work with the joint distribution $p_\theta(\mathbf{y},\mathbf{z})$ rather than with the marginal distribution $p_\theta(\mathbf{y})$ [33, Section 9.3]. Using Bayes' theorem, we can derive the posterior of the latent variables $p_\theta(\mathbf{z}|\mathbf{y})$ as:

$$p_\theta(\mathbf{z}|\mathbf{y}) = \frac{p_\theta(\mathbf{y},\mathbf{z})}{p_\theta(\mathbf{y})}. \tag{1.27}$$

$p_\theta(\mathbf{z}|\mathbf{y})$ can be seen as the reverse of the generative process $p_\theta(\mathbf{y}|\mathbf{z})$. It remains non-trivial to analytically compute $p_\theta(\mathbf{z}|\mathbf{y})$ due to the integral in the denominator. Given that $p_\theta(\mathbf{y},\mathbf{z})$ is relatively simple to compute, approximating one computationally intractable term in equation (1.27) may facilitate the evaluation of the other [153]. The following discusses how approximating the posterior distribution of the latent variables $p_\theta(\mathbf{z}|\mathbf{y})$ can assist in evaluating the marginal likelihood $p_\theta(\mathbf{y})$.

The true posterior distribution of latent variables $p_\theta(\mathbf{z}|\mathbf{y})$ can be approximated by a tractable variational distribution $q_\phi(\mathbf{z}|\mathbf{y})$, with ϕ denoting the variational parameters. To address the problem of approximating the true posterior, one can make use of variational inference, which measures the

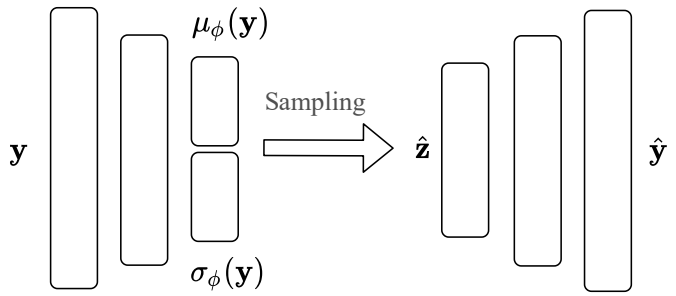


Fig. 1.3 Illustration of VAE with an encoder and decoder.

discrepancy of two distributions with the reverse KL divergence, i.e., $\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{y})||p_\theta(\mathbf{z}|\mathbf{y}))$ [112]. An easy-to-calculate approximate variational distribution is usually chosen, such as a commonly used Gaussian. For any choice of variational distribution, one can rewrite the approximation problem as:

$$\begin{aligned}\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{y})||p_\theta(\mathbf{z}|\mathbf{y})) &= \int q_\phi(\mathbf{z}|\mathbf{y}) \log \frac{q_\phi(\mathbf{z}|\mathbf{y})}{p_\theta(\mathbf{z}|\mathbf{y})} d\mathbf{z} \\ &= \log p_\theta(\mathbf{y}) - \mathcal{L}_{\phi,\theta}(\mathbf{y}),\end{aligned} \tag{1.28}$$

where $\mathcal{L}_{\phi,\theta}(\mathbf{y})$ is referred to as the ELBO and defined as

$$\mathcal{L}_{\phi,\theta}(\mathbf{y}) = \mathrm{E}_{q_\phi(\mathbf{z}|\mathbf{y})}[\log p_\theta(\mathbf{y}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{y})||p_\theta(\mathbf{z})). \tag{1.29}$$

The derivation can be found in, e.g., [33]. It indicates that minimizing the KL divergence between the approximate variational distribution $q_\phi(\mathbf{z}|\mathbf{y})$ and the true posterior distribution $p_\theta(\mathbf{z}|\mathbf{y})$ is equivalent to maximizing the ELBO. A simple re-arrangement of equation (1.28) gives:

$$\begin{aligned}\mathcal{L}_{\phi,\theta}(\mathbf{y}) &= \log p_\theta(\mathbf{y}) - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{y})||p_\theta(\mathbf{z}|\mathbf{y})) \\ &\leq \log p_\theta(\mathbf{y}).\end{aligned} \tag{1.30}$$

This indicates that the ELBO $\mathcal{L}_{\phi,\theta}(\mathbf{y})$ can act as a surrogate objective for optimizing the underlying distribution $\log p_\theta(\mathbf{y})$. In other words, maximizing the surrogate objective ELBO help reduce the gap to $\log p_\theta(\mathbf{y})$. When $\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{y})||p_\theta(\mathbf{z}|\mathbf{y})) = 0$, i.e., $q_\phi(\mathbf{z}|\mathbf{y})$ perfectly match the target $p_\theta(\mathbf{z}|\mathbf{y})$, the ELBO is equal to $\log p_\theta(\mathbf{y})$. Additionally, ELBO can also be derived using Jensen's inequality from the frequentist perspective [99].

The generative model VAE is trained by maximizing the ELBO. The VAE provides an efficient scheme to jointly model the approximate posterior $q_\phi(\mathbf{z}|\mathbf{y})$ and a stochastic data generator $p_\theta(\mathbf{y}|\mathbf{z})$. Specifically, the former is called the inference model (or recognition model), which is achieved by the probabilistic encoder, and the latter is called the generative model, which is achieved by the probabilistic decoder. ϕ and θ are comprised of the parameters of the encoder and decoder respectively. A VAE is illustrated in Figure 1.3.

Alternatively, the VAE can also be seen as a probabilistic version of autoencoders. Autoencoders refer to a particular type of DNN architecture, which features an encoder and decoder. The encoder network maps inputs into *deterministic* representations and the decoder network maps these latent representations back to the original data space. In contrast, the decoder of the VAE takes as input *probabilistic* representations $\mathbf{z}$ and can generate multiple valid target estimates by sampling in the latent space. Therefore, regularization is required to ensure a meaningful and constrained latent space. This is achieved through the KL divergence term in the optimization objective ELBO (1.29), which regularizes latent variables $\mathbf{z}$ to follow a prior distribution $p_\theta(\mathbf{z})$. Note that the prior distribution is usually chosen to be non-parameterized, such as a standard Gaussian. Eventually, training a VAE using stochastic gradient descent optimizes the encoder and decoder jointly, i.e., the parameters ϕ and θ [153].

Optimization of the ELBO using stochastic gradient techniques requires computing the gradient with respect to the parameters ϕ and θ, i.e., $\nabla_\phi \mathcal{L}_{\phi,\theta}(\mathbf{y})$ and $\nabla_\theta \mathcal{L}_{\phi,\theta}(\mathbf{y})$. This involves the expectation with respect to the variational distribution $q_\phi(\mathbf{z}|\mathbf{y})$. This becomes clearer when we expand the KL divergence in (1.29) and rearrange the ELBO into:

$$\begin{aligned}\mathcal{L}_{\phi,\theta}(\mathbf{y}) &= \mathrm{E}_{q_\phi(\mathbf{z}|\mathbf{y})}[\log p_\theta(\mathbf{y}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{y})||p_\theta(\mathbf{z})) \\ &= \mathrm{E}_{q_\phi(\mathbf{z}|\mathbf{y})}[\log p_\theta(\mathbf{y}|\mathbf{z})] - \mathrm{E}_{q_\phi(\mathbf{z}|\mathbf{y})}\left[\log\left(\frac{q_\phi(\mathbf{z}|\mathbf{y})}{p_\theta(\mathbf{z})}\right)\right] \\ &= \mathrm{E}_{q_\phi(\mathbf{z}|\mathbf{y})}\left[\underbrace{\log\frac{p_\theta(\mathbf{y},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{y})}}_{f(\mathbf{z})}\right].\end{aligned} \tag{1.31}$$

Since the variational distribution $q_\phi(\mathbf{z}|\mathbf{y})$ is not a function of θ, the gradient computation with respect to θ can be applied directly to the log-density ratio $f(\mathbf{z})$. However, this does not hold for variational parameters ϕ and therefore requires a stochastic estimator to obtain a tractable expectation [109, 154, 155]. To jointly train a neural network with parameters ϕ and θ, a reparameterization trick has been proposed for low-variance gradient approximation (where low variance can ensure fast convergences) [109, 154, 155]. The reparameterization technique assumes an invertible and differentiable transform, $\mathbf{z} = g_\phi(\varepsilon, \mathbf{y})$, where ε is another random variable. For example, when we have Gaussian-distributed latent variables $\mathcal{N}(\mu_\phi(\mathbf{y}), \sigma^2_\phi(\mathbf{y}))$, where $\mu_\phi(\mathbf{y})$ and $\sigma^2_\phi(\mathbf{y})$ denote the nonlinear mapping from the input to the mean and variance of the latent Gaussian variables, it can be standardized by:

$$\varepsilon = \frac{\mathbf{z} - \mu_\phi(\mathbf{y})}{\sigma_\phi(\mathbf{y})} \quad \text{with} \quad \varepsilon \sim p(\varepsilon), \tag{1.32}$$

where $p(\varepsilon)$ is a standard Gaussian with zero mean and unit variance. Then, the differential transform $g_\phi(\varepsilon, \mathbf{y})$ can be specified to be

$$\mathbf{z} = g_\phi(\varepsilon, \mathbf{y}) = \mu_\phi(\mathbf{y}) + \sigma_\phi(\mathbf{y})\varepsilon. \tag{1.33}$$

This reparameterization represents latent variables $\mathbf{z}$ as a function of another random variable ε, allowing transforming the expectation with respect to $q_\phi(\mathbf{y}|\mathbf{z})$ into the expectation with respect to $p(\varepsilon)$:

$$\mathrm{E}_{q_\phi(\mathbf{y}|\mathbf{z})}[f(\mathbf{z})] = \mathrm{E}_{p_{(\varepsilon)}}[f(g_\phi(\varepsilon,\mathbf{y}))]. \tag{1.34}$$

It can be seen that the standard Gaussian $p(\varepsilon)$ is not a function of ϕ and thus the gradient computation can be performed inside the expectation operator. Overall, the reparameterization gradient estimator allows for stable and efficient optimization by enabling direct computation of gradients with respect to the parameters of interest.

The VAE has been widely used to model the prior distribution of clean speech in speech signal processing [20, 21]. Several variants of the basic VAE framework have been developed [156–159]. For instance, besides utilizing Gaussian-distributed continuous latent variables, Kingma et al. [156] use additional categorically distributed discrete variables to incorporate class label information to form a conditional generation process. Similar ideas have also been applied to tasks involving speech modeling, where class labels can represent speech activity [160], phonemes [161], or speaker identities [161, 162]. Another conditional generative model to make diverse and structured predictions has been proposed by Sohn et al. [157]. Furthermore, Aksan et al. [158] proposed to capture temporal dependencies of input data through a combination of statistical hierarchical latent variables and temporal convolutional networks. Its excellent temporal modeling capabilities provide a good fit to handle speech data, which is inherently sequential. Richter [163] et al. applied it to speech enhancement and showed improved speech modeling capabilities than the vanilla VAE, which processes each frame independently. Similarly, Leglaive1 et al. [164] have also proposed a probabilistic generative model based on recurrent neural networks to better model the temporal activity of speech signals. Girin et al. [165] presented a consolidated overview of several variants of the VAE with a focus on modeling temporal dependencies within sequential data. Their following work in [22] has also experimentally evaluated the performance of these variants on the speech enhancement task.

Studies on the VAE have also focused on disentangling latent variables from speech data and incorporating other modalities such as visual information. For example, Hsu et al. [159] leveraged the observation that some attributes of speech vary slightly within an utterance but significantly across utterances, while others vary similarly both within and between utterances, to effectively disentangle sequence-level and segment-level attributes of speech signals. Also, a similar idea to disentangle global speaker information and local content representation has been studied in [166]. Carbajal et al. [167] proposed to disentangle speech activity from other latent variables through adversarial training, which is achieved in combination with visual modality. Indeed, despite the effectiveness of the VAE-based speech model, it is still challenging to estimate the parameters of clean speech from noisy mixtures in challenging noise conditions, such as low SNRs. For this, audio-visual speech enhancement can incorporate complementary features from visual information, which is independent of acoustic interference. Sadeghi et al. followed the conditional generative framework [157] and extended the audio-only VAE [20] to incorporate visual information for speech

enhancement, specifically lip movements [168]. A further improvement was proposed in [169] to process audio and video modalities with separate encoders. Nguyen et al. applied a similar idea to audio-visual speech separation [170]. Their subsequent extensions [171, 172] tackled the cases of unreliable visual inputs by switching between audio-only and audio-visual VAE-based priors.

Existing research has sought to improve speech modeling by exploring various methods that take into account different characteristics of speech signals. As part of the thesis, we present a method [P6] in Chapter 3 to further refine speech modeling by incorporating noise information while at the same time retaining its generative capabilities. Another advantage is that the proposed method can complement the current advances in speech modeling. This is because it maintains the same architecture as the VAE model presented in [20], making it easy to combine with other techniques.

1.4.2 Deep Generative Speech Enhancement

As discussed in the previous sections, representation learning techniques, such as the VAE and NMF, have found their widespread use in speech and audio signal processing. The adaptable nature of lightweight NMF models makes them suitable for various problem settings. Meanwhile, the data-driven nature of VAE models enables them to learn prior distributions from complex, high-dimensional data. Recent research has been conducted to integrate the VAE and NMF into a unified Bayesian framework, combining the benefits of both techniques [21, 20]. Specifically, it has been proposed to combine a VAE-based speech model and a NMF-based noise model, showing superior results than the algorithms based exclusively on NMF in speech enhancement [20, 38]. Aiming to achieve high robustness to unseen acoustic conditions, the NMF-based noise model is not pre-trained but adapted to individual samples during testing, that is, the noise parameters are estimated directly from input noisy mixtures. Therefore, only the parameters of the VAE are learned on isolated clean speech data during training, making the algorithm semi-supervised. Such methods have also demonstrated improved generalization capabilities over supervised masking baselines [21, 129].

Given a pre-trained VAE-based speech model and an untrained NMF-based noise model, various Bayesian optimization methods have been proposed to perform speech enhancement. These optimization strategies may define stochastic latent variables and deterministic unknown parameters differently [62]. For example, besides speech latent variables in the VAE, Bando et al. [21] adopted a fully Bayesian model and placed conjugate prior distributions on the NMF parameters, independently on each element in the dictionary and activation matrices. At inference, the posterior distributions of the random variables are sampled iteratively using an MCMC algorithm. A similar optimization method has also been developed for multichannel speech enhancement, where in addition to speech latent variables and stochastic elements in the dictionary and activation matrices, a prior distribution is also imposed on the spatial covariance matrix [173]. In contrast, Leglaive et al. [20] proposed to perform maximum likelihood estimation using an MCEM optimization method, where only speech latent representations in the VAE are considered as random variables, and the dictionary and activation matrices in NMF are seen as unknown parameters. The posterior distribution of speech latent vari-

ables is evaluated using a Metropolis-Hastings sampling algorithm, and unknown NMF parameters are updated by deriving multiplicative update rules with auxiliary function techniques [41, 42]. A further extension to the multichannel case is presented in [38], where the additional spatial covariance matrix is seen as an unknown parameter and updated by solving the Riccati equations as in [10]. This optimization strategy has been applied or adapted in several studies [160, 167, 163, 168, 170]. Besides, different variants of variational EM algorithms have been presented, such as [164, 174, 171]. For instance, Pariente et al. [174] derived a variational EM algorithm by treating the speech latent variables and the complex coefficients of speech and noise as a set of latent variables. This is followed by further factorizing the joint posterior distribution of the latent variables using mean-field variational inference principles [33]. Leglaive et al. formulated the optimization problem using a fixed form variational inference strategy in [164], where the variational distribution of speech latent variables is predefined as a Gaussian parameterized by the encoder of the VAE at both training and inference. At inference, the encoder of the VAE is further tuned based on the noisy input using gradient-based optimization techniques, while the unknown NMF parameters are updated similar to [20]. Other optimization techniques involve reusing and adapting the decoder of the VAE include, e.g., [129, 175].

In this thesis, we follow the MCEM optimization methods [20, 38], which can yield good approximation results given sufficient computational power. Furthermore, our contribution in the thesis is in principle independent of specific optimization strategies and thus has the potential to be extended to many existing optimization formulations. Next, we introduce the VAE-NMF framework and the MCEM optimization strategy presented in [20] in detail.

NMF-Based Noise Model and VAE-Based Speech Model

Complex spectral coefficients of noise are independently modeled using a NMF-based Gaussian:

$$N_{ft} \sim \mathcal{N}_{\mathbb{C}}(N_{ft}; 0, [\mathbf{WH}]_{ft}). \tag{1.35}$$

The VAE has been combined with the local Gaussian assumption [20, 176] to model the complex spectral coefficients of clean speech. Specifically, the decoder, parameterized by $\boldsymbol{\theta}$, models the spectral variance conditioned on the latent variable $\mathbf{z}$. The generative model can then be represented by:

$$S_{ft}|\mathbf{z}_t \sim \mathcal{N}_{\mathbb{C}}(S_{ft}; 0, \sigma_f^2(\mathbf{z}_t)), \tag{1.36}$$

where $\mathbf{z}_t \in \mathbb{R}^L$ denotes a latent random vector at t-th frame and $\sigma_f^2(\mathbf{z}_t) : \mathbb{R}^L \rightarrow \mathbb{R}_+$ denotes the nonlinear mapping function generating the speech tempo-spectral power conditioned on the latent variables. The inference model, i.e., the VAE encoder, infers latent variables $\mathbf{z}_t$ from the observed data $\mathbf{s}_t = \{S_{1t}, \cdots, S_{Ft}\} \in \mathbb{C}^F$, which is a vector of complex speech coefficients at the t-th time frame. It is typically assumed that the posterior of latent variables $q_{\boldsymbol{\phi}}(\mathbf{z}_t|\mathbf{s}_t)$ follows a real-valued Gaussian distribution

$$\mathbf{z}_t|\mathbf{s}_t \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}}(|\mathbf{s}_t|^2),\ \boldsymbol{\sigma}_{\mathbf{z}}(|\mathbf{s}_t|^2)), \tag{1.37}$$

where $\mu_{\mathbf{z}}(|\mathbf{s}_t|^2) : \mathbb{R}_+^F \to \mathbb{R}^L$ and $\sigma_{\mathbf{z}}(|\mathbf{s}_t|^2) : \mathbb{R}_+^F \to \mathbb{R}_+^L$ denote the nonlinear mapping from the power spectrogram to the mean and variance of the latent variables. Finally, the parameters of the encoder and decoder can be jointly optimized by maximizing the ELBO

$$\mathcal{L}_{\phi,\theta}(\mathbf{s}_t) = \mathrm{E}_{q_\phi(\mathbf{z}_t|\mathbf{s}_t)}[\log p_\theta(\mathbf{s}_t|\mathbf{z}_t)] - \mathrm{KL}(q_\phi(\mathbf{z}_t|\mathbf{s}_t)||p(\mathbf{z}_t)), \tag{1.38}$$

where $p(\mathbf{z}_t)$ represents the standard Gaussian prior of $\mathbf{z}_t$. The first term is a reconstruction term, which measures how effectively the decoder can reconstruct the input from the latent representation. The second is the regularization term, which ensures that the latent variable follows a prior distribution, promoting a continuous latent space.

The reconstruction term in the ELBO involves computing the expectation of $\log p_\theta(\mathbf{s}_t|\mathbf{z}_t)$ with respect to the posterior $q_\phi(\mathbf{z}_t|\mathbf{s}_t)$, which can not be solved analytically; therefore it is approximated by a MC estimator with R samples. It has been pointed out that in practice, when the batch size is large enough, the number of samples R can be set to 1 [109]. Furthermore, the posterior of latent variables is factorized as the product of elements in the vector, i.e., $q_\phi(\mathbf{z}_t|\mathbf{s}_t) = \prod_{l=1}^L q_\phi(Z_{tl}|\mathbf{s}_t)$ and $\mathbf{z}_t = \{Z_{t1}, \cdots, Z_{tL}\}$. With these assumptions, the ELBO (1.38) is approximated by:

$$\mathcal{L}_{\phi,\theta}(\mathbf{s}_t) \approx -\frac{1}{R}\sum_{r=1}^{R}\sum_{f=1}^{F} \mathrm{IS}(|S_{ft}|^2, \sigma_f^2(\mathbf{z}_t^{(r)})) + \frac{1}{2}\sum_{l=1}^{L}\left[\ln \sigma_{\mathbf{z},l}^2(|\mathbf{s}_t|^2) - \mu_{\mathbf{z},l}(|\mathbf{s}_t|^2)^2 - \sigma_{\mathbf{z},l}^2(|\mathbf{s}_t|^2)\right], \tag{1.39}$$

where $\mathbf{z}_t^{(r)}$ denotes the r-th sample from the posterior and $\mathrm{IS}(\cdot,\cdot)$ denotes the Itakura-Saito divergence.

Eventually, with an additive model, the noisy mixture signal can be written as:

$$X_{ft} = \sqrt{g_t} S_{ft} + N_{ft}, \tag{1.40}$$

where $g_t \in \mathbb{R}_+$ is a gain parameter to increase the robustness to the time-varying loudness of speech sounds. Given the VAE-based speech model and the NMF-based noise model, under the independence assumption of speech and noise, the noisy mixture coefficients X_{ft} follow:

$$X_{ft} \sim \mathcal{N}_{\mathbb{C}}(0,\ g_t \sigma_f^2(\mathbf{z}_t) + [\mathbf{WH}]_{ft}). \tag{1.41}$$

Next, we will introduce the MCEM optimization strategy to solve the model with latent variables and unknown parameters.

Monte Carlo expectation-maximization algorithm

For probability models involving latent variables and unknown parameters, the expectation-maximization algorithm is often used to find maximum likelihood solutions [33]. It is performed in a two-step iterative manner: *expectation (E)* and *maximization (M)*. For example, one can first derive the lower bound of a log-likelihood [33, Sec. 9.4]. In step E, we fix the unknown parameters to their current values and maximize the lower bound with respect to the variational distribution introduced. Given

the current values of unknown parameters, the solution is obtained when the lower bound is equal to the log-likelihood. This is given by the posterior distribution of latent variables given the currently fixed parameters. In the M step, the variational distribution of latent variables is fixed to the solution obtained in the preceding E step, and the lower bound is maximized with respect to the unknown parameters [177]. The function being maximized at this step is the expectation of the complete-data log-likelihood [33].

We observe only noisy mixtures X_{ft} during testing, from which we need to infer proper random variables $\mathbf{z}_t$ and unknown parameters $\zeta = \{g_t, [\mathbf{WH}]_{ft}\}$, as shown in (1.41). A solution to this additive model can be derived by following the iterative optimization technique to approximate the likelihood function [33, 177]. The complete-data log-likelihood is denoted by

$$\log p_\zeta(\mathbf{x},\mathbf{z}) = \sum_{f=1}^{F}\sum_{t=1}^{T} \log p_\zeta(\mathbf{x}_t,\mathbf{z}_t) \overset{c.}{=} -\sum_{f=1}^{F}\sum_{t=1}^{T} \log\left(g_t\sigma_f^2(\mathbf{z}_t) + [\mathbf{WH}]_{ft}\right) + \frac{|X_{ft}|^2}{g_t\sigma_f^2(\mathbf{z}_t) + [\mathbf{WH}]_{ft}}, \tag{1.42}$$

where $\overset{c.}{=}$ denotes equality up to a constant. A set of the mixture STFT coefficients is denoted by $\mathbf{x} = \{\mathbf{x}_1, \cdots, \mathbf{x}_t, \cdots \mathbf{x}_T\}$ and $\mathbf{x}_t = \{X_{1t}, \cdots, X_{ft}, \cdots, X_{Ft}\}$. Let $\mathbf{z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_t, \cdots \mathbf{z}_T\}$ denote a set of latent vectors. The complete-data log-likelihood incorporates both speech and noise models, and the dependency on the VAE-based speech model is omitted for notation simplicity since its parameters remain fixed during inference. We follow an MCEM algorithm [20] in this thesis. Note that during inference, we can only access noisy mixtures rather than the posterior of latent variables $q_\phi(\mathbf{z}_t|\mathbf{s}_t)$ in (1.38). Therefore, the speech latent variables $\mathbf{z}$ are inferred from the noisy mixtures, i.e. $p_\zeta(\mathbf{z}_t|\mathbf{x}_t)$,

In the **E** step, the current values of the parameter set are denoted as ζ^*, which we use to evaluate the posterior distribution $p_{\zeta^*}(\mathbf{z}|\mathbf{x})$. The posterior distribution is then used to compute the expectation of the complete-data log-likelihood, denoted as $\mathcal{Q}(\zeta, \zeta^*)$

$$\mathcal{Q}(\zeta,\zeta^*) = \mathrm{E}_{p_{\zeta^*}(\mathbf{z}|\mathbf{x})}\left[\log p_\zeta(\mathbf{x},\mathbf{z})\right]. \tag{1.43}$$

Since the posterior $p_{\zeta^*}(\mathbf{z}|\mathbf{x})$ can not be computed analytically, we need to perform approximate inference. Here, a MCMC method is used, specifically, a Metropolis-Hastings (MH) algorithm with a symmetric Gaussian as the proposal distribution. The MH method constructs a Markov chain with the posterior $p_{\zeta^*}(\mathbf{z}_t|\mathbf{x}_t)$ as its stationary distribution for all frames $t \in [1, \cdots, T]$. When sampling iteratively, the sample at the m-th iteration is drawn from the proposal random distribution constructed based on the preceding step [20]:

$$\mathbf{z}_t|\mathbf{z}_t^{(m-1)} \sim \mathcal{N}(\mathbf{z}_t^{(m-1)}, \varepsilon^2\mathbf{I}), \tag{1.44}$$

where ε^2 is a hyperparameter. The samples are selectively accepted according to the probability:

$$\alpha_t = \min\left(1,\ \frac{p_{\zeta^*}(\mathbf{x}_t|\mathbf{z}_t)p(\mathbf{z}_t)}{p_{\zeta^*}(\mathbf{x}_t|\mathbf{z}_t^{(m-1)})p(\mathbf{z}_t^{(m-1)})}\right), \tag{1.45}$$

where the independence assumption is often applied to $\mathbf{x}_t$ such that $p_{\zeta^*}(\mathbf{x}_t|\mathbf{z}_t)$ can be written as the product of the elements in the t-th frame vector [20]. When ideally, the sampling process arrives at the stationary distribution, one can approximate the expectation in (1.43) with a MC estimator by sampling from the posterior R times:

$$\mathcal{Q}(\zeta, \zeta^*) \approx \frac{1}{R} \sum_{r=1}^{R} \log p_\zeta(\mathbf{x}, \mathbf{z}^{(r)}). \tag{1.46}$$

In the **M** step, the function $\mathcal{Q}(\zeta, \zeta^*)$ is maximized with respect to all parameters in the set ζ:

$$\widetilde{\zeta} = \arg\max_{\zeta} \mathcal{Q}(\zeta, \zeta^*), \tag{1.47}$$

where $\widetilde{\zeta}$ denotes a new set of parameters after the optimization. Here, the update rules for each parameter in ζ can be derived using a block-coordinate approach [42]. With the complex Gaussian assumptions of speech and noise, the function $\mathcal{Q}(\zeta, \zeta^*)$ can be decomposed into a superposition of a convex term and a concave term, which can be bounded using the auxiliary function techniques. For example, a convex function can be bounded using Jensen's inequality, while a concave function can be bounded using a first-order Taylor expansion [42, 20, 38]. Consequently, the optimization of the function $\mathcal{Q}(\zeta, \zeta^*)$ turns into the optimization of an upper bound, and the multiplicative update rules for the parameters in ζ are derived accordingly.

Given the estimates of unknown parameters, the speech spectral coefficients are obtained by computing the speech posterior mean [20]:

$$\widehat{S}_{ft} = \mathrm{E}_{p_\zeta(\mathbf{z}_t|\mathbf{x}_t)} \left[\frac{g_t \sigma_f^2(\mathbf{z}_t)}{g_t \sigma_f^2(\mathbf{z}_t) + [\mathbf{WH}]_{ft}} \right] X_{ft}, \tag{1.48}$$

where $\widehat{S}_{ft}$ is an estimate of the speech spectral coefficients. The expectation can be approximated by drawing samples using the Metropolis-Hastings algorithm similar to the E step.

Extension to Multichannel VAE-NMF

By applying the additive model to the multi-channel mixtures, we have:

$$\mathtt{X}_{ft} = \mathtt{S}_{ft} + \mathtt{N}_{ft}, \tag{1.49}$$

where $\mathtt{X}_{ft}$, $\mathtt{S}_{ft}$, and $\mathtt{N}_{ft} \in \mathbb{C}^M$ represent the multichannel spectral coefficients of the noisy speech, clean speech, and noise, respectively. The VAE-NMF can be extended to the multichannel scenarios accordingly, where additional spatial information can be exploited. We model the speech and noise spectral coefficients with multivariate complex Gaussian distributions:

$$\mathtt{S}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \sigma_f^2(\mathbf{z}_t)\boldsymbol{R}_{\mathtt{S},f}), \quad \text{and} \quad \mathtt{N}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, [\mathbf{WH}]_{ft}\boldsymbol{R}_{\mathtt{N},f}), \tag{1.50}$$

where $\boldsymbol{R}_{\mathrm{S},f} \in \mathbb{C}^{M\times M}$ and $\boldsymbol{R}_{\mathrm{N},f} \in \mathbb{C}^{M\times M}$ are the spatial covariance matrices of speech and noise, respectively. The temporal-spectral power of speech and noise are modeled by the VAE and NMF respectively, i.e., by $\sigma_f^2(\mathbf{z}_t)$ and $(\mathbf{WH})_{ft}$. We can pre-train the VAE model on clean speech at the training stage and estimate the parameters of the NMF-based noise model on the noisy mixtures using a MCEM algorithm presented in [38]. Compared to the single-channel scenario, the multichannel VAE-NMF also requires estimating the spatial covariance matrices of speech and noise, which is achieved by solving the Riccati equations as in [10]. Finally, the spectral coefficients of clean speech can be obtained through multichannel Wiener filtering [38].

As previously discussed, the parameters of NMF are estimated based on noisy mixtures at the inference stage. This allows the VAE-NMF method to potentially adapt to various acoustic conditions. However, it remains a challenge to perform robustly on challenging acoustic scenarios, such as those involving non-stationary noise. In this thesis, we investigate its application in human-robot interaction scenarios, which involves dealing with not only environmental noise but also ego-noise. To increase the algorithm's robustness to noise, we take advantage of the less diverse nature of ego-noise and extend the NMF-based noise model to incorporate this prior information. The proposed scheme has been applied to both single-channel [P7] and multichannel [P8] applications, resulting in improved overall speech enhancement performance.

1.5 Thesis Outline

Predictive approaches are trained to learn mappings from input features to output targets. However, their estimation mechanisms are difficult to interpret due to the black-box nature of DNNs. Although supervised deep predictive models have been the dominant tool in speech enhancement, they often exhibit limited generalization ability under complex and unseen acoustic conditions. In this thesis, we endeavor to overcome these limitations by exploring uncertainty modeling and developing methods based on statistically principled generative frameworks. Essentially, we aim to incorporate statistical modeling into deep speech enhancement to improve its interpretability and robustness.

1. Uncertainty in Deep Predictive Speech Enhancement

Research Questions	
RQ1	How can uncertainty be effectively modeled in deep predictive speech enhancement, and to what extent can uncertainty estimates reliably predict deviations from ground-truth speech? How does uncertainty estimation affect speech enhancement performance?
RQ2	How can one leverage statistical domain knowledge to develop more efficient methods for uncertainty estimation in deep predictive speech enhancement?
RQ3	Can uncertainty estimates be further leveraged to improve the robustness and generalization ability of speech enhancement systems?

The first part of this thesis is centered around predictive time-frequency masking, which has demonstrated remarkable speech enhancement performance in existing work. It can leverage non-linear modeling capabilities provided by DNNs to approximate various training targets. Despite being widely used, predictive methods still face challenges in understanding how particular solutions are derived from training data. In particular, these approaches are often framed as single-output estimation problems, and their estimates are accepted without questioning their accuracy. As a result, these approaches may result in fundamentally incorrect estimates for unseen samples without any indication that these estimates are uncertain. This naturally gives rise to uncertainty in the model's predictions. Modeling uncertainty allows us to quantify and interpret the predictive confidence, especially when the network model is processing out-of-distribution samples. Various sources in a DNN framework can contribute to uncertainty in the predictions. While uncertainty modeling has received increasing attention in computer vision and deep learning, its study in DNN-based speech enhancement remains under-explored.

The thesis explores the uncertainty in predictions arising from various aspects of deep speech enhancement. The first question we aim to answer concerns *how uncertainty can be modeled and captured in deep predictive speech enhancement and how uncertainty estimation impacts speech enhancement performance.* For this, we start with the Gaussian priors for the speech and noise spectral coefficients and use a neural network to estimate the full clean speech posterior distribution. This statistical assumption is further incorporated into Bayesian deep learning frameworks. Based on this, we present a comprehensive analysis of uncertainty estimates on a time-frequency bin scale and evaluate the uncertainty-augmented speech enhancement methods over different datasets. Second, we observe that generic task-agnostic uncertainty modeling methods may involve a costly sampling process, which poses challenges for applications under resource-constrained conditions. Thus, we delve into the question of *whether domain-specific statistical knowledge allows for efficient uncertainty estimation with negligible computational overhead*. We assume a complex Gaussian mixture model for speech and noise and employ a DNN to estimate the resulting speech posterior distribution, which also has the form of a CGMM. With this, we demonstrate the potential to obtain the uncertainty in

predictions arising from both data uncertainty and model uncertainty with only a single forward pass of DNNs. Uncertainty can help assess the reliability of a model's predictions by quantifying the degree of deviation from the true data. This motivates the research question of *how uncertainty estimates can be leveraged to improve the robustness and generalization ability of speech enhancement systems.* We explore its application in two tasks: unsupervised domain adaptation for speech enhancement and modality fusion in audio-visual phoneme recognition. For domain adaptation, we develop an uncertainty-based filtering strategy that selects high-quality speech estimates from an unlabeled target domain to fine-tune a model trained on the source domain. For multimodal information fusion, we investigate how uncertainty arising from noisy audio and video modalities can lead to a more informed fusion scheme in audio-visual phoneme recognition.

2. Noise-Aware Generative Speech Enhancement Based on Variational Autoencoder and Non-Negative Matrix Factorization

Research Questions	
RQ4	How can VAE-NMF-based generative approaches leverage prior noise knowledge to improve speech modeling capabilities for better generalization in unseen acoustic environments?
RQ5	Can one derive a flexible and effective noise adaptation scheme that can reuse learned noise representation while adapting to unseen noise characteristics? Furthermore, can such an adaptation scheme be extended to multichannel applications?

To address the interpretability and generalization issues, we depart from predictive principles and focus instead on deep generative speech enhancement in the latter half of the thesis. Deep generative models aim to learn prior distributions of given data and reuse the learned knowledge to perform speech enhancement. A recent framework built on generative VAE and NMF [21, 20] models provides an elegant combination of DNNs and statistical models, demonstrating a promising generalization ability to unseen acoustic conditions. More specifically, a VAE is used to learn a prior distribution of clean speech, which is then combined with an untrained NMF-based noise model. The parameters of the VAE can be obtained beforehand by training on isolated clean speech, while the NMF parameters can be adapted to individual noisy samples during testing. Thus, it leverages the benefits of data-driven methods, while the adaptability to test inputs ensures its generalization ability to diverse noise conditions. Notably, this framework is statistically principled and the unknown parameters can be derived in some optimal statistical sense. However, training the VAE only on clean speech data makes the algorithm susceptible to noise presence during testing. As a result, the VAE-NMF framework lacks noise-robustness in acoustically challenging conditions, such as when clean speech is distorted simultaneously by non-stationary ego-noise generated by interactive robots and unseen background noise [178].

The second part of the thesis focuses on introducing noise information into deep generative approaches and extending their application to challenging human-robot-interaction scenarios, which involve both ego-noise and environmental noise. When dealing with ego-noise distortion, we leverage the limited-degree-of-freedom nature of ego-noise and augment the noise model accordingly by incorporating this prior knowledge. When multichannel ego-noise recordings are available, we are able to additionally make use of spatial information of noise sources distributed over the robot's body. In this way, we aim to leverage structured characteristics of ego-noise, spectrally and/or spatially, to gain robustness to ego-noise distortions, while at the same time retaining the ability to adapt to unseen environmental noise. We demonstrate the potential advantages by comparing the resulting noise adaptation scheme to the baselines that either learn noise on-the-fly or rely solely on learned prior knowledge. In addition to improving noise modeling capabilities, we also want to improve the robustness of the VAE-based speech model. To this end, we employ paired noisy-clean speech data during training to improve the inference of the latent variables of clean speech, leading to a noise-aware encoder for the VAE. Thus, the core research topic in this part centers on *how the speech and noise models can be improved by incorporating prior noise knowledge while maintaining the statistical interpretability of the framework.*

1.6 Related Publications

This cumulative thesis is based on the following publications. We categorized the publications into two groups according to the two main research topics we explored. We include [P2], [P3], [P4], [P5], [P6], [P7], and [P8], in the main part of the thesis and enclose them with gray boxes. The conference publication [P1] serves as the basis for the extended journal publication [P2]. Therefore, there is an overlap in content, and [P1] is included in Appendix A.

Chapter 2 Uncertainty in Deep Predictive Speech Enhancement

2.1 Integrating Uncertainty into Neural Network-Based Speech Enhancement

[P1] H. Fang, T. Peer, S. Wermter, and T. Gerkmann, "Integrating statistical uncertainty into neural network-based speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Singapore, Singapore, 2022, pp. 386–390.

Contribution: The experimental results are obtained by Fang. Fang wrote the paper. Peer provided feedback through discussions about the results, helped create the method diagram, and reviewed the final manuscript. Gerkmann provided feedback on the ideas and intermediate results and reviewed the final manuscript. Wermter contributed to providing feedback on the final manuscript.

[P2] H. Fang, D. Becker, S. Wermter, and T. Gerkmann, "Integrating uncertainty into neural network-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 31, pp. 1587–1600, 2023.

Contribution: The experimental results are obtained by Fang. Fang wrote the paper. Gerkmann provided feedback on the ideas and intermediate results and reviewed the final manuscript. Becker and Wermter contributed to providing feedback on the final manuscript.

2.2 Uncertainty Estimation in Deep Speech Enhancement Using Complex Gaussian Mixture Models

[P3] H. Fang and T. Gerkmann, "Uncertainty estimation in deep speech enhancement using complex Gaussian mixture models," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.

Contribution: The experimental results are obtained by Fang. Fang wrote the paper. Gerkmann provided feedback on the ideas and intermediate results and reviewed the final manuscript.

2.3 Uncertainty-Based Remixing for Unsupervised Domain Adaptation in Deep Speech Enhancement

[P4] H. Fang, and T. Gerkmann, "Uncertainty-based remixing for unsupervised domain adaptation in deep speech enhancement," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aalborg, Denmark, 2024, pp. 45-49.

Contribution: The experimental results are obtained by Fang. Fang wrote the paper. Gerkmann provided feedback on the ideas and intermediate results and reviewed the final manuscript.

2.4 Uncertainty-Driven Hybrid Fusion for Audio-Visual Phoneme Recognition

[P5] H. Fang, S. Frintrop, and T. Gerkmann, "Uncertainty-driven hybrid fusion for audio-visual phoneme recognition," in *Speech Communication; 15th ITG Conference*, Aachen, Germany, 2023, pp. 255–259.

Contribution: The experimental results are obtained by Fang. Fang wrote the paper. Gerkmann provided feedback on the ideas and intermediate results and reviewed the final manuscript. Frintrop contributed to providing feedback on the final manuscript.

Chapter 3 Noise-Aware Generative Speech Enhancement Based on Variational Autoencoder and Non-Negative Matrix Factorization

3.1 Variational Autoencoder for Speech Enhancement with a Noise-Aware Encoder

[P6] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, ON, Canada, 2021, pp. 676–680.

Contribution: The experimental results are obtained by Fang. Fang wrote the paper. Carbajal and Gerkmann provided feedback on the ideas and intermediate results and reviewed the final manuscript. Wermter contributed to providing feedback on the final manuscript.

3.2 Joint Reduction of Ego-Noise and Environmental Noise with a Partially-Adaptive Dictionary

[P7] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Joint reduction of ego-noise and environmental noise with a partially-adaptive dictionary," in *Speech Communication; 14th ITG Conference*, online, 2021, pp. 1–5.

Contribution: The experimental results are obtained by Fang. Fang wrote the paper. Carbajal and Gerkmann provided feedback on the ideas and intermediate results and reviewed the final manuscript. Wermter contributed to providing feedback on the final manuscript.

3.3 Partially Adaptive Multichannel Joint Reduction of Ego-Noise and Environmental Noise

[P8] H. Fang, N. Wittmer, J. Twiefel, S. Wermter, and T. Gerkmann, "Partially adaptive multichannel joint reduction of ego-noise and environmental noise," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.

Contribution: The experimental results are based on the master thesis of Wittmer supervised by Fang. Fang wrote the paper. Gerkmann provided feedback on the ideas and intermediate results and reviewed the final manuscript. Twiefel and Wermter contributed to providing feedback on the final manuscript.

CHAPTER 2

Uncertainty in Deep Predictive Speech Enhancement

2.1 Integrating Uncertainty into Neural Network-Based Speech Enhancement [P2]

Reference

H. Fang, D. Becker, S. Wermter, and T. Gerkmann, "Integrating uncertainty into neural network-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 31, pp. 1587–1600, 2023. DOI: 10.1109/TASLP.2023.3265202

Copyright Notice

Integrating Uncertainty into Neural Network-based Speech Enhancement

Huajian Fang*†, *Student Member, IEEE*, Dennis Becker†,
Stefan Wermter†, *Member, IEEE*, and Timo Gerkmann* , *Senior Member, IEEE*

***Abstract*—Supervised masking approaches in the time-frequency domain aim to employ deep neural networks to estimate a multiplicative mask to extract clean speech. This leads to a single estimate for each input without any guarantees or measures of reliability. In this paper, we study the benefits of modeling uncertainty in clean speech estimation. Prediction uncertainty is typically categorized into *aleatoric uncertainty* and *epistemic uncertainty*. The former refers to inherent randomness in data, while the latter describes uncertainty in the model parameters. In this work, we propose a framework to jointly model aleatoric and epistemic uncertainties in neural network-based speech enhancement. The proposed approach captures aleatoric uncertainty by estimating the statistical moments of the speech posterior distribution and explicitly incorporates the uncertainty estimate to further improve clean speech estimation. For epistemic uncertainty, we investigate two Bayesian deep learning approaches: Monte Carlo dropout and Deep ensembles to quantify the uncertainty of the neural network parameters. Our analyses show that the proposed framework promotes capturing practical and reliable uncertainty, while combining different sources of uncertainties yields more reliable predictive uncertainty estimates. Furthermore, we demonstrate the benefits of modeling uncertainty on speech enhancement performance by evaluating the framework on different datasets, exhibiting notable improvement over comparable models that fail to account for uncertainty.**

***Index Terms*—Speech enhancement, Bayesian estimator, uncertainty estimation, deep neural networks**

I. Introduction

Speech recorded in noisy environments is often corrupted by background noise, which renders it difficult to understand by either humans or machines via automatic speech recognition systems. These problems call for robust speech enhancement algorithms, which extract desired clean speech from noisy mixtures to improve speech quality and intelligibility of recordings. In this paper, we consider single-channel speech enhancement.

Speech enhancement algorithms typically utilize the short-time Fourier transform (STFT) to transfer the recorded signal into the time-frequency domain, where multiplicative filters can be applied to obtain an estimate of clean speech [1], [2]. Various Bayesian estimators, e.g., maximum a posteriori (MAP) and minimum mean squared error (MMSE) estimators, have been developed based on different statistical distributions about speech and noise, aiming to restore either the spectral coefficients of the STFT or the spectral magnitudes [3]–[6]. Given the assumption that speech is degraded by uncorrelated additive noise and that both follow complex Gaussian distributions with zero mean, the well-known Wiener filter can be derived. Traditionally, the speech and noise variances estimated by statistical model-based methods [1], [7] can be used to construct the MMSE-optimal Wiener filter.

Recently, neural networks have been widely used in speech enhancement methods due to their flexibility and effectiveness in nonlinear modeling. Depending on their application, varying degrees of success are reported [8]–[19]. Specifically, deep neural networks have been utilized to replace some of the building blocks of conventional speech enhancement methods. For instance, a neural network-based speech presence probability estimator has been proposed in [8] and combined with a single-channel multi-frame approach [9]. In [10], [11], neural networks are employed to estimate speech and noise power spectrum densities that are required in various Bayesian estimators. Additionally, recent work has leveraged the probabilistic modeling of generative networks for speech enhancement. For example, the variational autoencoder (VAE) has been used to estimate the clean speech distribution, which is then combined with a separate noise model to construct a noise reduction Wiener filter [12], [14]. The robustness of this filter can be further improved by injecting noise information [16], temporal dependencies [20]–[22], and information from other modalities, such as vision [17], [23]. Besides, speech enhancement approaches based on perceptual metric-guided adversarial training [24], [25] and diffusion-based generative models [26], [27] have also been presented. In contrast, supervised masking approaches [18] aim to learn a mapping from the noisy input to a masking filter. It allows neural networks to directly estimate a time-frequency filter by training on a large amount of noisy-clean speech pairs using an appropriate cost function [19]. In this work, we focus on supervised masking approaches.

While the time-frequency noise-removing filter aims to remove noise with minimum speech distortions, the algorithm's robustness and reliability are not guaranteed, especially when speech is corrupted by previously unobserved noise. To alleviate this shortcoming, research has been conducted to investigate how to generalize to unseen situations by, e.g., developing more sophisticated network architectures, improved features, or including more training data that covers a wide variety of acoustic scenarios [28]–[30]. The first is often accompanied by a tremendous increase in model parameters, while the latter is rather time-demanding. Still, improving the generalization ability of neural networks in unseen scenarios is an unsolved problem considering the black-box nature of neural networks. It is thus necessary and beneficial to obtain the associated uncertainty as an indicator of reliability besides the point estimate, especially when the model is processing out-of-distribution samples that are insufficiently represented by training data.

In machine learning, predictive uncertainty is typically decom-

The authors gratefully acknowledge support from the German Research Foundation DFG under the project CML (TRR 169) and ahoi.digital.

The authors are with the *Signal Processing (SP) Group, the †Knowledge Technology (WTM) Group, Department of Informatics, Universität Hamburg, 22527 Hamburg, Germany (e-mail:{huajian.fang; dennis.becker-1; stefan.wermter; timo.gerkmann}@uni-hamburg.de).

posed into two categories [31]–[33]: *aleatoric* uncertainty and *epistemic* uncertainty. The term aleatoric uncertainty is used to describe the uncertainty of an estimate due to the intrinsic randomness of noisy observations. For speech enhancement, it originates from the stochastic nature of both speech and noise and is reflected in the variance of the clean speech posterior predictive distribution. Epistemic uncertainty is of different nature: If the parameters of a neural network are trained, e.g., using different training data, different initialization, or a different number of epochs, different parameters result. Therefore, also the parameters of a neural network used to estimate clean speech are uncertain. This uncertainty of the parameters is called epistemic uncertainty (also known as *model uncertainty*). For a general introduction to uncertainty modeling, readers are suggested to refer to a review article by Hüllermeier et al. [31]. Various uncertainty measures have been employed in the deep regression setting, such as confidence intervals, differential entropy, and variance. Depeweg et al. [34] propose to measure uncertainty based on the entropy of the predictive distribution, which represents the information level of random variables. Pearce et al. [35] use confidence intervals (which state how certain the estimate is within a certain range) in a distribution-free setting. In this paper, we address uncertainty modeling in a probabilistic way following [33], [36], [37] and measure the uncertainty in terms of the *variance*.

Aleatoric uncertainty. Due to the stochastic nature of speech and noise, a mapping from noisy speech to clean speech is uncertain as reflected by the posterior predictive distribution of clean speech. We can model this posterior using a specific conditional distribution, such as a Gaussian or a Laplacian [33], [36], [38], and employ a neural network to directly estimate the statistical moments of this distribution. While the predicted mean is the MMSE estimate of the target [2], the associated variance can be used to quantify the data inherent uncertainty, i.e., aleatoric uncertainty [33].

Few studies in neural network-based speech enhancement have incorporated the uncertainty of aleatoric nature. Chai et al. propose to use a generalized Gaussian distribution to model the prediction error on a logarithmic scale [39]. In [40], a neural network is used to estimate the parameters of a Gaussian mixture model, which then serves as the basis of an extra statistical model-based speech enhancement approach. This results in only a slight improvement over the baseline optimized with the MMSE criterion. Siniscalchi [41] leverages neural networks to learn a histogram distribution to approximate the conditional target speech distribution, which is assumed to be a truncated Gaussian distribution with a fixed variance in each frequency band. However, the fixed variance does not help to capture data-dependent uncertainty.

Epistemic uncertainty. Estimating the statistical moments of the speech posterior predictive distribution allows capturing aleatoric uncertainty, but fails to account for epistemic uncertainty, which corresponds to the uncertainty in neural network parameters [31]–[33]. Epistemic uncertainty can be captured using Bayesian inference approaches, which instead of modeling the parameters of a neural network as *deterministic* values, place a distribution over the network parameters and estimates the posterior distribution of the *stochastic* network parameters [33]. By sampling from the posterior network parameter distribution, multiple sets of neural network parameter realizations can be obtained, thus producing multiple output predictions for each input sample. Uncertainty in predictions due to epistemic uncertainty can be empirically quantified by the variance in these output predictions [31], [33]. While the true posterior network distribution is intractable [42], it can be approximated using 1) Markov Chain Monte Carlo (MCMC) methods [43], [44], which are sampling-based approaches that construct a Markov Chain with the posterior network parameter distribution as its stationary distribution, 2) variational inference [42], [45], [46], which approximates the true posterior network parameter distribution with a tractable variational distribution, and 3) ensemble approaches [36], [47], [48], which were proposed from the frequentist perspective but are considered as an approximate Bayesian approach [37], [49]. For instance, Gal et al. [42] perform variational inference and interpret the dropout regularization technique [50] as imposing Bernoulli distributions on the neural network's weights. This method, referred to as Monte Carlo dropout (*MC dropout*), provides a set of target estimates from multiple forward passes by activating dropout at inference. This set of predictions can empirically approximate the outcome distribution for each input sample and allows inference of the variance (i.e., epistemic uncertainty). In contrast, *Deep ensembles* proposed in [36] can quantify epistemic uncertainty by training multiple neural networks with random weight initialization [37], [38].

Recent studies attempt to consider the uncertainty of epistemic nature in, e.g., speech emotion recognition [51], [52] and speech recognition [53]–[55]. In [51], epistemic uncertainty is captured in a speech emotion recognition model for selective prediction, where samples with low confidence (high uncertainty) are rejected. Braun et al. [53] apply a Gaussian distribution to the weights of an end-to-end speech recognition model to capture uncertainty of neural network parameters, which is then used for parameter pruning. In a recent publication [54], epistemic uncertainty is employed to improve the robustness of domain adaptation for speech recognition. However, quantifying epistemic uncertainty in neural network-based speech enhancement remains unexplored.

Contributions. Capturing overall predictive uncertainty, which reflects both aleatoric and epistemic uncertainties, is challenging, especially for deep neural networks, but crucial for an understanding of the model's prediction behaviour. In this work, we propose a method that allows capturing aleatoric uncertainty and combining it with epistemic uncertainty approximations to quantify overall predictive uncertainty. In the context of neural network-based speech enhancement, to the best of our knowledge, this is the first work to study different sources of uncertainty in a joint framework and provides for systematic analyses.

We follow the complex Gaussian speech-plus-noise assumption and propose to train a neural network to estimate the Wiener filter and its variance, which quantifies aleatoic uncertainty, based on the MAP inference of *complex spectral coefficients*. To regularize the variance estimation, we build an approximate MAP (AMAP) estimator of *spectral magnitudes* using the estimated Wiener filter (mean of the complex clean speech posterior predictive distribution) and uncertainty (variance of the complex clean speech posterior distribution) explicitly. The resulting AMAP estimator is in turn used in conjunction with the MAP inference of complex spectral coefficients to form a novel hybrid loss function. Rather than discarding uncertainty information at inference, the proposed scheme allows us to explicitly incorporate aleatoric uncertainty approximations into clean speech estimation in a principled way to further correct erroneous speech estimates.

Previous studies on modeling epistemic uncertainty have

focused on other tasks than speech enhancement, e.g., [38], [51]–[56]. Yet, questions such as how reliable and accurate the estimates of epistemic uncertainty are in speech enhancement, and how modeling epistemic uncertainty affects enhancement performance, have not been addressed. To this end, we investigate two Bayesian deep learning techniques: MC dropout [42] and Deep ensembles [36] to capture epistemic uncertainty in clean speech estimation due to their efficiency in approximating Bayesian inference. Although previous works have explored ensemble-based speech enhancement methods [57], [58], they did not investigate the effectiveness of ensemble-based methods for uncertainty estimation.

Moreover, we propose to estimate overall predictive uncertainty reflecting both aleatoric and epistemic uncertainties by combining the proposed hybrid loss function with the ensemble-based method. Finally, we present a comprehensive analysis of uncertainty from different sources and show their impacts on speech enhancement performance over different datasets, which we hope lays the foundation for further use of uncertainties.

This paper extends our previous conference publication [59], which studied aleatoric uncertainty. Here, we propose to additionally capture epistemic uncertainty and combine them to quantify overall predictive uncertainty in clean speech estimation. Furthermore, we provide a more detailed analysis with respect to uncertainty estimates from different sources in a joint framework. Section II describes the signal model. In Section III, we propose to estimate the uncertainty of aleatoric nature following the complex Gaussian-distributed speech posterior and present how this uncertainty can be incorporated into clean speech estimation. In Section IV, we show how to capture epistemic uncertainty and quantify overall predictive uncertainty that combines different sources of uncertainty. We introduce the experimental setting in Section V, analyze uncertainty estimates in Section VI, and present enhancement performance in Section VII. Section VIII summarizes the findings.

II. Signal Model

In the single-channel speech enhancement problem, the noisy mixture consists of clean speech and additive noise. We apply the STFT to obtain the representation in the time-frequency domain as:

$$X_{ft} = S_{ft} + N_{ft}, \tag{1}$$

where X_{ft}, S_{ft}, and N_{ft} represent the complex spectral coefficients of mixture, speech, and noise, at the time frame $t \in \{1,2,\cdots,T\}$ and the frequency bin $f \in \{1,2,\cdots,F\}$. T and F denote the number of time frames and frequency bins respectively. The objective is to recover clean speech in the time-frequency domain by applying a multiplicative filter. To derive such a filter, various assumptions are made according to different signal characteristics. By assuming that the speech and noise coefficients are uncorrelated and follow a circularly symmetric complex Gaussian distribution,

$$S_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2_{s,ft}), \quad N_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2_{n,ft}), \tag{2}$$

where $\sigma^2_{s,ft}$ and $\sigma^2_{n,ft}$ represent the variances of speech and noise respectively, the likelihood $p(X_{ft}|S_{ft})$ follows a complex Gaussian distribution with mean S_{ft} and variance $\sigma^2_{n,ft}$, given by

$$p(X_{ft}|S_{ft}) = \frac{1}{\pi \sigma^2_{n,ft}} \exp\left(-\frac{|X_{ft} - S_{ft}|^2}{\sigma^2_{n,ft}} \right). \tag{3}$$

With the likelihood in (3) and the prior in (2), we can apply Bayes' theorem to obtain the posterior distribution of clean speech as a complex Gaussian of the form [2]:

$$p(S_{ft}|X_{ft}) = \frac{1}{\pi \lambda_{ft}} \exp\left(-\frac{|S_{ft} - W^{\text{WF}}_{ft} X_{ft}|^2}{\lambda_{ft}} \right), \tag{4}$$

$$W^{\text{WF}}_{ft} = \frac{\sigma^2_{s,ft}}{\sigma^2_{s,ft} + \sigma^2_{n,ft}}, \quad \lambda_{ft} = \frac{\sigma^2_{s,ft} \sigma^2_{n,ft}}{\sigma^2_{s,ft} + \sigma^2_{n,ft}}. \tag{5}$$

W^{WF}_{ft} is recognized as the *Wiener filter* and λ_{ft} is the variance of the posterior distribution. Under this assumption, the MMSE estimator, which corresponds to the expectation of the posterior distribution, leads to the Wiener filter applied as:

$$\widetilde{S}_{ft} = W^{\text{WF}}_{ft} \cdot X_{ft}. \tag{6}$$

Due to the symmetry of the complex Gaussian distribution, the MAP estimator of complex speech coefficients is identical to the MMSE estimator.

III. Aleatoric Uncertainty Estimation

Although speech enhancement is typically formulated as a problem with a single output, the dependency between input and output can be modeled stochastically by means of a speech posterior predictive distribution $p(S_{ft}|X_{ft})$, i.e., a variance λ_{ft} is associated with the clean speech estimate and can be interpreted as a measure of uncertainty of the Wiener estimate [2]. This uncertainty accounts for random effects in data and is referred to as *aleatoric uncertainty* [33], [36]. When properly captured, aleatoric uncertainty can reflect the expected estimation error in the absence of ground truth.

A. Deep Aleatoric Uncertainty Estimation

In contrast to traditional signal processing techniques [1], [2], [60], where the Wiener filter is constructed by separately estimating the variances of speech and noise from the noisy mixture X_{ft}, neural network-based supervised masking methods allow direct estimation of multiplicative filters. Besides the Wiener filter W^{WF}_{ft}, one can further estimate the data-dependent aleatoric uncertainty λ_{ft} if the neural network is optimized using the speech posterior predictive distribution (4), i.e., by minimizing the negative logarithm of the posterior distribution of clean speech $p(S_{ft}|X_{ft})$ (the logarithm does not affect the optimization problem due to monotonicity) and averaging over time-frequency bins:

$$\widetilde{W}^{\text{WF}}_{ft}, \widetilde{\lambda}_{ft} = \underset{W^{\text{WF}}_{ft}, \lambda_{ft}}{\operatorname{argmin}} \underbrace{\frac{1}{FT} \sum_{f,t} \log(\lambda_{ft}) + \frac{|S_{ft} - W^{\text{WF}}_{ft} X_{ft}|^2}{\lambda_{ft}}}_{\mathcal{L}_{p(S|X)}}, \tag{7}$$

where $\widetilde{W}^{\text{WF}}_{ft}$, $\widetilde{\lambda}_{ft}$ denote estimates of the Wiener filter and associated aleatoric uncertainty [33], [36].

In contrast, if we assume a constant uncertainty for all time-frequency bins, i.e., $\lambda_{ft} = \lambda^*$, and refrain from explicitly optimizing for λ^*, $\mathcal{L}_{p(S|X)}$ degenerates into the well-known mean squared error (MSE) loss

$$\mathcal{L}_{\text{MSE}} = \frac{1}{FT} \sum_{f,t} |S_{ft} - W^{\text{WF}}_{ft} X_{ft}|^2, \tag{8}$$

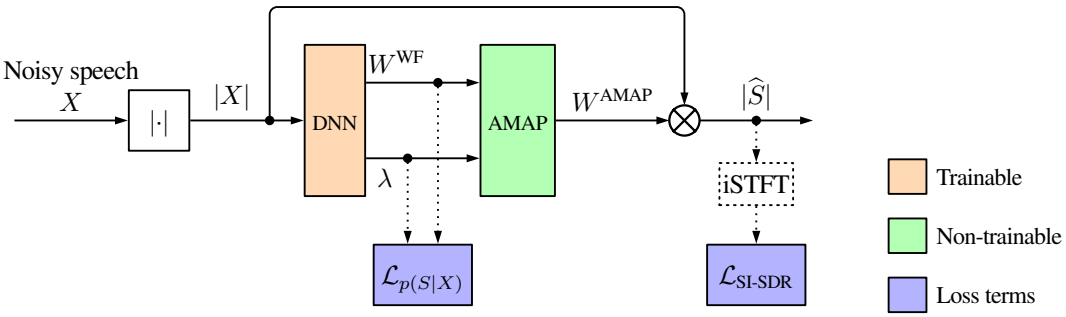


Fig. 1: Block diagram of the proposed neural network-based aleatoric uncertainty estimation.

which is widely used in neural network-based regression tasks including speech enhancement [19]. However, neural networks trained to perform point estimation do not necessarily output reliable estimates for clean speech when processing out-of-distribution samples that are underrepresented by the training data [28]. In this work, we discard the assumption of constant uncertainty; instead, we propose to treat uncertainty estimation as an additional task by training a neural network with the negative log speech posterior $\mathcal{L}_{p(S|X)}$. Consequently, this method not only allows us to obtain a noise-removing mask, but also empowers the model to capture the uncertainty of aleatoric nature associated with predictions.

Modeling aleatoric uncertainty by minimizing the logarithm of the posterior predictive distribution results in an improvement over baselines that fail to account for uncertainty in computer vision tasks [33]. However, directly using $\mathcal{L}_{p(S|X)}$ as the loss function is prone to overfitting [59] and may result in reduced estimation performance of the Wiener filter. A recent publication [61] also reveals that directly minimizing the logarithm of the conditional probability hinders the training of mean estimation, which leads to premature convergence. To tackle this problem, we propose an additional regularization of the loss function by incorporating the estimated uncertainty into clean speech estimation as described next.

B. Joint Enhancement and Uncertainty Estimation

Estimating uncertainty λ_{ft} associated with the Wiener filter is challenging since ground truth of uncertainty is not readily available. Instead, uncertainty estimation is an unsupervised task with an unspecified search space, which can potentially lead to unstable training [62], [63]. In this work, we propose to incorporate a subsequent speech enhancement task that explicitly uses both the Wiener filter and its uncertainty λ_{ft} during the training procedure. The speech enhancement task provides additional coupling between the outputs (Wiener filter and uncertainty). In this manner, the neural network is guided to estimate the uncertainty values that are relevant to the speech enhancement task, as well as to enhance the estimation of the Wiener filter.

Considering complex coefficients with a symmetric posterior (4), the MAP and MMSE estimators both lead directly to the Wiener filter W_{ft}^{WF} and do not require an uncertainty estimate. However, this situation changes if we consider spectral magnitude estimation. The magnitude posterior $p(|S_{ft}||X_{ft})$, derived by integrating the phase out of (4), follows a Rician distribution [4]

$$p(|S_{ft}||X_{ft}) = \frac{2|S_{ft}|}{\lambda_{ft}} \exp\left(-\frac{|S_{ft}|^2+(W_{ft}^{\text{WF}})^2|X_{ft}|^2}{\lambda_{ft}}\right) I_0\left(\frac{2|X_{ft}||S_{ft}|W_{ft}^{\text{WF}}}{\lambda_{ft}}\right), \quad (9)$$

where $I_0(\cdot)$ is the modified zeroth-order Bessel function of the first kind.

In order to compute the MAP estimate for the spectral magnitude, the mode of the Rician distribution has to be estimated, which is difficult to do analytically. However, it can be approximated by substituting a Bessel function approximation following [64] into (9) and maximizing with respect to the spectral magnitude, yielding a simple closed-form expression [2], [4]:

$$\begin{aligned} |\widehat{S}_{ft}| &\approx W_{ft}^{\text{AMAP}}|X_{ft}| \\ &= \left(\frac{1}{2}W_{ft}^{\text{WF}} + \sqrt{\left(\frac{1}{2}W_{ft}^{\text{WF}}\right)^2 + \frac{\lambda_{ft}}{4|X_{ft}|^2}}\right)|X_{ft}|, \end{aligned} \quad (10)$$

where $|\widehat{S}_{ft}|$ is an estimate of the clean spectral magnitude $|S_{ft}|$ using the AMAP estimator of spectral magnitudes W_{ft}^{AMAP}. It can be noticed that the estimator W_{ft}^{AMAP} utilizes both the Wiener filter W_{ft}^{WF} and the associated uncertainty λ_{ft}. Fig. 2 illustrates the input-output estimation characteristics of the AMAP estimator and Wiener filter [2]. We can see that W_{ft}^{AMAP} is *nonlinear* with respect to the noisy input and tends to cause less target attenuation than the Wiener filter especially for low inputs. This indicates that incorporating the associated uncertainty λ_{ft} may increase the robustness of the estimator by potentially preserving more speech at the slight cost of noise removal.

After combining the estimated magnitude $|\widehat{S}_{ft}|$ with the noisy phase, we can apply the inverse STFT to obtain an estimate of the time-domain speech signal, denoted as $\widehat{s}$. Afterwards, the estimated time-domain signal is used to compute the negative scale-invariant

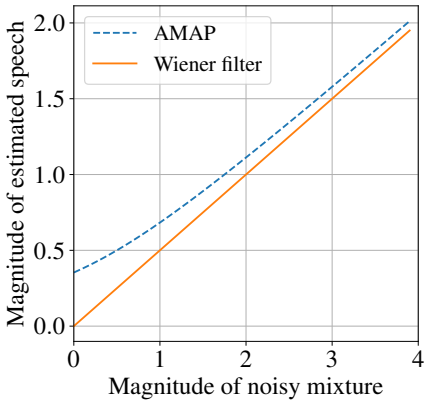


Fig. 2: Input-output characteristics of the AMAP estimator W_{ft}^{AMAP} and Wiener filter W_{ft}^{WF} (setting $\sigma_{s,ft}^2 = \sigma_{n,ft}^2 = 1$ in this example).

signal-to-distortion ratio (SI-SDR) metric [65]:

$$\mathcal{L}_{\text{SI-SDR}} = -10\log_{10}\left(\frac{||\alpha s||^2}{||\alpha s - \widehat{s}||^2}\right), \quad \alpha = \frac{\widehat{s}^T s}{||s||^2}, \tag{11}$$

which is leveraged as an additional term in the loss function that forces the speech estimate (computed with W_{ft}^{AMAP}) to be similar to the time-domain clean speech target s. While a spectrum loss like (8) is a straightforward solution to regularize the uncertainty estimation, the time-domain loss is expected to be more effective since it is directly related to the raw waveform, implicitly taking phase information into account and thus promoting speech reconstruction for better perceptual performance [66].

Eventually, we propose to combine the SI-SDR loss $\mathcal{L}_{\text{SI-SDR}}$ with the negative log-posterior $\mathcal{L}_{p(S|X)}$ given in (7), and train the neural network using a hybrid loss function

$$\mathcal{L} = \beta\mathcal{L}_{p(S|X)} + (1-\beta)\mathcal{L}_{\text{SI-SDR}}, \tag{12}$$

with the weighting factor $\beta \in [0,1]$. By explicitly using the estimated uncertainty for the speech enhancement task, the hybrid loss guides both mean and variance estimation to improve speech enhancement performance. Fig. 1 depicts a block diagram of this approach.

IV. Bayesian uncertainty estimation

While neural networks performing point estimation have demonstrated effectiveness in speech enhancement, it is not guaranteed that neural networks can generalize well to unfamiliar acoustic situations. Therefore, to quantify the overall predictive confidence regarding the estimated clean speech, it is necessary to also assess the uncertainty of the neural network parameters (i.e., *epistemic uncertainty*). Note that a single neural network optimized using the proposed hybrid loss (12) allows capturing aleatoric uncertainty but is unaware of epistemic uncertainty. To solve this, we can utilize Bayesian deep learning approaches, assuming that the weights of a neural network follow some probability distribution rather than deterministic values. Furthermore, when combined with the loss (12), an ensemble of networks can provide both aleatoric uncertainty and epistemic uncertainty estimates.

A. Epistemic Uncertainty Estimation

Bayesian deep learning provides a set of principled methods to capture epistemic uncertainty [36], [42]–[44], [46], [48]. Early work on MCMC methods [43], [44] constructs a Markov chain with the posterior network parameter distribution as its stationary distribution and generates multiple network parameter realizations by sampling from this distribution. However, MCMC methods are computationally inefficient and do not scale well to neural networks with a large number of parameters [37], [38]. Recent work based on variational inference allows approximating the true posterior network parameter distribution with a tractable distribution [45], [46], while at the same time ensemble-based methods are proposed as simple and scalable frequentist alternatives to model uncertainty [36], [47], [48]. Among the existing Bayesian deep learning methods, MC dropout and Deep ensembles have shown their scalability in large neural network-based problems, such as semantic segmentation [33] and depth estimation [37]. Here, we investigate their effectiveness for uncertainty estimation in speech enhancement.

We define a neural network as a function parameterized by θ and a training dataset that contains noisy-clean speech pairs $\mathcal{D} = \{(S_{11},X_{11}),...,(S_{FT},X_{FT})\}$. Hereafter we omit the indices ft, since all time-frequency bins are treated independently in (4). Since the posterior network parameter distribution $p(\theta|\mathcal{D})$ is computationally intractable in a high dimensional space, variational inference approximates the true posterior network parameter distribution by a pre-specified variational distribution $q(\theta)$ and the speech posterior predictive distribution at inference time is obtained by marginalizing out $q(\theta)$ as:

$$\begin{aligned} p(S|X,\mathcal{D}) &= \int p(S|X,\theta)p(\theta|\mathcal{D})d\theta \\ &\approx \frac{1}{M}\sum_{m=1}^{M} p(S|X,\theta_m), \quad \theta_m \sim q(\theta), \end{aligned} \tag{13}$$

where θ_m represents m-th sampling from $q(\theta)$ [67]. MC dropout approximates the posterior network parameter distribution using the Bernoulli distribution and samples neural network weights by activating dropout at inference time. Gal et al. provide further details on the derivations in [42]. This allows obtaining M target speech estimates from multiple stochastic forward passes for each input. In contrast, Deep ensembles repeatedly train the same model M times with random initialization and random data shuffling [36], generating M neural networks with deterministic network parameter estimates $\{\theta_m\}_{m=1}^{m=M}$. Since θ_m can be viewed as independent samples from a certain approximate distribution $q(\theta)$, Deep ensembles can be considered equivalent to approximate Bayesian inference [37]. Therefore, the predictive distribution is obtained similarly to (13). Furthermore, neural networks usually contain a large number of parameters, which makes them multi-modal in the parameter space. Different initialization starting points in Deep ensembles allow the neural network to converge to different local optima, thus potentially capturing multiple modes of $p(\theta|\mathcal{D})$ [37], [48].

Epistemic uncertainty can be approximated by building an ensemble of neural networks using either MC dropout or Deep ensembles, where each network is trained to estimate the Wiener filter only with the loss function $\mathcal{L}_{\text{MSE}}$ (8). With the results of M forward passes, we can approximate the mean and variance of the distribution $p(S|X)$ by the empirical mean and variance of the prediction set [38], [42]:

$$\widetilde{S} = \frac{1}{M}\sum_{m=1}^{M}\widetilde{S}_{\theta_m}, \quad \widetilde{\Sigma} = \frac{1}{M}\sum_{m=1}^{M}|\widetilde{S}_{\theta_m} - \widetilde{S}|^2, \tag{14}$$

where $\widetilde{S}_{\theta_m}$ denotes clean speech estimated using the neural network with parameters θ_m. $\widetilde{S}$ represents the average clean speech estimate and $\widetilde{\Sigma}$ quantifies the epistemic uncertainty.

B. Overall Predictive Uncertainty

In the case of optimizing the network using (12), besides the Wiener estimate $\widetilde{S}_{\theta_m}$, each neural network with weights θ_m can produce the associated variance $\widetilde{\lambda}_{\theta_m}$. The overall predictive uncertainty, which reflects both aleatoric and epistemic uncertainties, can be computed using the law of total variance [33], [38]:

$$\widetilde{S} = \frac{1}{M}\sum_{m=1}^{M}\widetilde{S}_{\theta_m}, \quad \widehat{\Sigma} = \frac{1}{M}\sum_{m=1}^{M}\left(|\widetilde{S}_{\theta_m} - \widetilde{S}|^2 + \widetilde{\lambda}_{\theta_m}\right), \tag{15}$$

where $\widetilde{S}$ denotes the average Wiener estimate, and $\widehat{\Sigma}$ quantifies the overall predictive uncertainty.

For each neural network with weights θ_m, we can further generate the AMAP clean speech estimate $\widehat{S}_{\theta_m}$ by explicitly incorporating the associated uncertainty $\widetilde{\lambda}_{\theta_m}$ as in (10). Therefore, given an ensemble of networks, besides the average Wiener estimate $\widetilde{S}$, the average AMAP estimate can be obtained by:

$$\widehat{S} = \frac{1}{M}\sum_{m=1}^{M} \widehat{S}_{\theta_m}. \tag{16}$$

V. Experimental setting

A. Datasets

For training and validation, we use a subset of the Deep Noise Suppression (DNS) Challenge's training set [68], which contains synthetic audio samples of 100 hours with signal-to-noise ratios (SNRs) uniformly distributed between -5 dB and 20 dB. The dataset is randomly split into 80 and 20 hours for training and validation respectively. The model is evaluated on two different unseen datasets. The first is the reverb-free synthetic test set released by DNS Challenge. This evaluation dataset is disjoint from the training and validation datasets and is created by adding noise signals sampled from 12 categories [68] to speech signals from [69] at SNRs distributed between 0 dB and 25 dB [68]. The second unseen evaluation dataset is created using clean speech from the evaluation subset of WSJ0 (`si_et_05`) [70] and four types of noise from CHiME3 (`cafe`, `street`, `pedestrian`, and `bus`) [71]. The SNRs are randomly selected from {-10 dB, -5 dB, 0 dB, 5 dB, 10 dB}.

B. Architecture and Hyperparameters

To ensure a fair comparison, all experiments are performed based on the same U-Net neural network architecture [72], [73]. The U-Net structure with skip connections between the encoder and the decoder is comprised of several blocks, each of which consists of: 2D convolution layer + instance normalization [74] + Leaky ReLU with slope 0.2. The encoder contains 6 blocks that increase the feature channel from 1 to 512 progressively ($1-16-32-64-128-256-512$), while the decoder reduces it back to 16 ($512-256-128-64-32-16-16$), followed by a 1×1 convolution layer that outputs a mask of the same shape as the input. For all blocks, the kernel size is set to $(5,5)$ with stride $(1,2)$ and padding $(2,2)$, processing a 2-D input with a dimension of (T,F). For the model estimating aleatoric uncertainty, the output layer is split into two heads that predict both the Wiener filter and associated uncertainty[1]. We applied the sigmoid activation function to the estimated Wiener filter, while using the *log-exp* technique to constrain the uncertainty output to be greater than 0, i.e., the network outputs the logarithm of the variance, which is then recovered by the exponential term in the loss function. The batch size is 64; the learning rate is 0.001; the weight decay parameter is set to 0.0005. All neural networks are trained with the Adam optimizer [75]. The training process is stopped if the validation loss fails to decrease for 10 consecutive epochs and the learning rate is halved when the validation loss does not decrease for 3 epochs.

[1]Code for the model is available at: https://github.com/sp-uhh/uncertainty-SE.

The noisy-clean speech pairs have a sampling rate of 16 kHz, and the STFT is computed using a 32 ms Hann window with 50% overlap.

C. Methods

The algorithms considered in this work include:

1) *Baseline WF*: The U-Net architecture was trained on noisy-clean speech pairs using loss function (8). This serves as a baseline, assuming a constant variance for all time-frequency bins and estimating the Wiener filter for each input only.
2) *Baseline SI-SDR*: Following the same constant variance assumption as *Baseline WF*, the U-Net network was trained to output a multiplicative filter and optimized using the time-domain loss function (11). This serves as another baseline that fails to account for uncertainty.
3) *Aleatoric-WF & Aleatoric-AMAP*: The hybrid loss function (12) allows us to generate two possible clean estimates for each input, i.e., by using the estimated Wiener filter (6) or by applying the AMAP estimator (10) that incorporates both the Wiener filter and its associated uncertainty. They are denoted as *Aleatoric-WF* and *Aleatoric-AMAP* respectively. We observe experimentally that the performance of Aleatoric-AMAP only fluctuates slightly with different β values, while the performance of Aleatoric-WF decreases when the value of β is large. The weighting factor β was empirically chosen to be 0.001 to achieve a good trade-off between the performance of Aletoric-WF and Aleaotirc-AMAP.
4) *MC dropout*: Inserting dropout after each convolution layer regularizes too strongly and impacts the model performance [56], which was confirmed in our preliminary experiments. We thus studied several variants of the U-Net by inserting the dropout layer at different positions of the architecture, and selected the variant with three dropout layers (drop probability of 0.5 [37], [50], [56]) inserted after the three deepest blocks of the encoder. The same cost function as *Baseline WF* is used. This method captures epistemic uncertainty by activating the dropout layers at inference.
5) *Deep ensembles*: The same setup as *Baseline WF* was trained M times with random initialization. This allows the model to capture epistemic uncertainty.
6) *DE-Aleatoric-WF & DE-Aleatoric-AMAP*: The same setup as *Aleatoric-WF/AMAP* was trained M times with random intialiation. This allows capturing aleatoric and epistemic uncertainties simultaneously. We average over the estimates according to (15) and (16) to obtain two clean speech estimates: *DE-Aleatoric-WF* and *DE-Aleatoric-AMAP* respectively.

VI. Analysis of uncertainty estimation

In this section, we introduce the evaluation metrics for uncertainty and then analyze the captured aleatoric and epistemic uncertainties. Finally, we show that combining two types of uncertainty yields more reliable predictive uncertainty.

A. Uncertainty Evaluation Metrics

To evaluate the captured uncertainty, the sparsification plot and the sparsification error are used as evaluation metrics [37], [38],

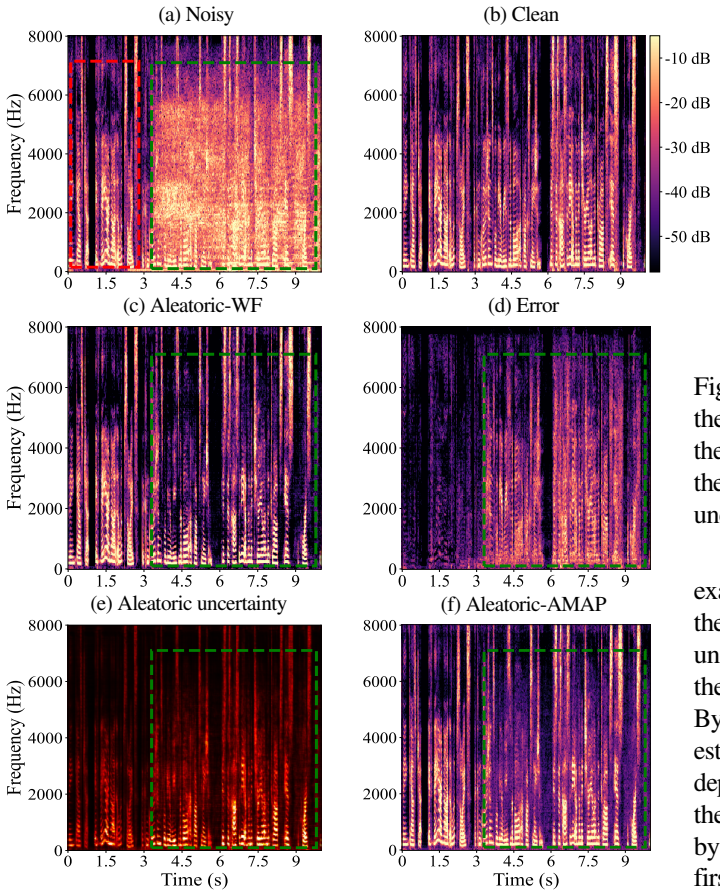


Fig. 3: Aleatoric uncertainty (shown in (e)) captured by the proposed loss function (12) for an excerpt from the DNS test dataset. The uncertainty is visualized as a heatmap. The black color indicates low uncertainty, whereas the brighter color indicates higher uncertainty.

[76]. The sparsification plot illustrates the correlation between the uncertainty measure and the true error. The error of a time-frequency bin is defined as the absolute square between the estimated spectral coefficient and the ground-truth. For this plot, the errors in the time-frequency domain are first sorted according to their corresponding uncertainty measures. The residual error should gradually decrease when the time-frequency bins with large uncertainties are removed. This leads to a plot of the root mean squared error (RMSE) versus the fraction of removed time-frequency bins. Normalization is applied to ensure that the plot is initialized at 1. The best ordering of uncertainty measures is determined by ranking the true errors [38], [76]. This provides a lower bound of each sparsification plot, denoted as the *oracle* curve, i.e., when the uncertainty estimates and errors are perfectly correlated, the sparsification plot and the oracle curve coincide. The sparsficiation error is computed as the difference between the sparsification plot and the corresponding oracle curve, and the area under the sparsification error (AUSE) curve provides a single value that enables comparison of different uncertainty modeling techniques. A lower AUSE value (i.e., the closer the sparsification plot is to its oracle curve) indicates a more accurate estimate of uncertainty.

B. Analysis of Aleatoric Uncertainty Estimation

In this part, we analyze the captured data-dependent aleatoric uncertainty associated with the Wiener estimate. For this, an audio

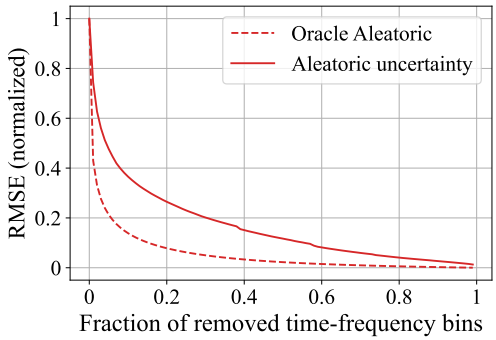


Fig. 4: Sparsification plot of aleatoric uncertainty $\tilde{\lambda}$ evaluated on the DNS test dataset. The dashed line denotes the lower bound of the sparsification plot of aleatoric uncertainty. A smaller distance of the sparsification plot to the oracle curve indicates a more accurate uncertainty estimation.

example from the DNS challenge test set is selected to illustrate the effectiveness of the proposed optimization metric in modeling uncertainty. Aleatoric-WF in Fig. 3 (c) shows the spectrogram of the clean speech obtained by applying the estimated Wiener filter. By computing the absolute square between the clean reference and estimated spectral coefficients, we can obtain the estimation error as depicted in Fig. 3 (d). It can be observed that large errors occur when the speech is heavily disturbed by noise, as in the region marked by the green box, while for inputs with less distortion, such as the first three seconds, the model produces smaller errors. Meanwhile, the proposed loss function enables the estimation of uncertainty associated with the Wiener filter, as shown in Fig. 3 (e), denoted as aleatoric uncertainty. It shows that aleatoric uncertainty prevails in speech presence regions. By relating Fig. 3 (d) to Fig. 3 (e), the model outputs relatively large uncertainty (e.g., the green box-marked part) when large errors are produced. This suggests that the neural network is able to produce reasonable uncertainty estimates when dealing with complex unseen inputs. Furthermore, we can incorporate the estimated uncertainty into clean speech inference, as in (10), which leads to a clean speech estimate shown in Fig. 3 (f), denoted as Aleatoric-AMAP. It is observed that more speech is preserved than Aleatoric-WF in the highly-uncertain green box-marked region at some cost of noise reduction, i.e., Aleatoric-AMAP leads to less speech distortion with a slight tendency of retaining more noise. The reason for this is that with reliable uncertainty estimates, Aleatoric-AMAP can increase the estimator's value in (10) under high uncertainty (as the AMAP estimator's value is positively correlated with the uncertainty estimate when other terms are fixed), thus causing less target attenuation.

Besides the qualitative analysis, we can associate the captured uncertainty with the corresponding prediction errors on the time-frequency bin scale and use sparsification plots to analyze the reliability of the uncertainty estimates. The sparsification plot shown in Fig. 4 is computed based on all audio samples in the DNS reverb-free test dataset. We observe a rapid decrease at the beginning in Fig. 4, implying that large errors come with large uncertainty estimates. By removing 20 percent of time-frequency bins with high uncertainty (i.e., 0.2 in the horizontal axis), the RMSE value drops by around two-thirds. Thus, the monotonically decreasing sparsification plot in Fig. 4 again suggests that the predicted aleatoric

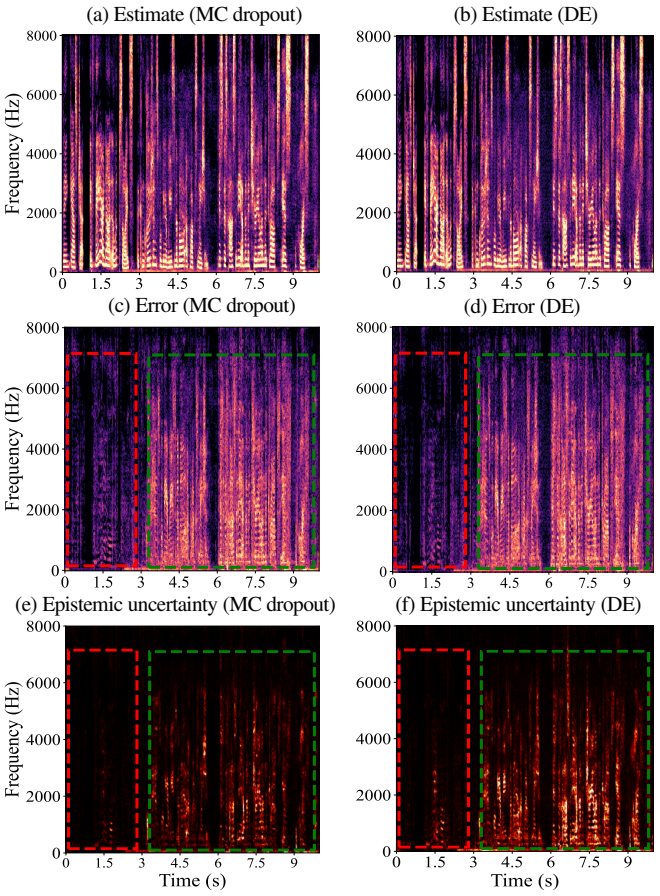


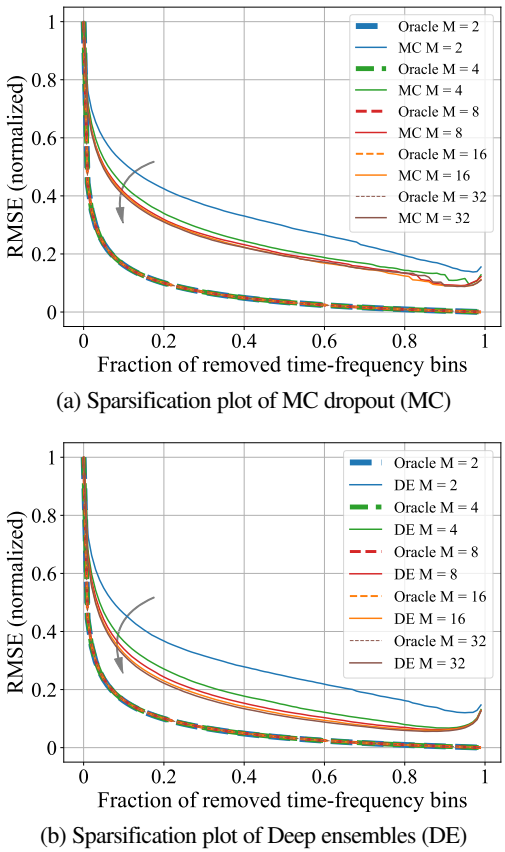
Fig. 5: The same excerpt as in Fig. 3 illustrates the captured epistemic uncertainty obtained by applying Bayesian deep learning methods ($M = 16$). *Estimate (MC dropout)* and *Estimate (DE)* represent clean speech estimated using MC dropout and Deep ensembles.

uncertainty measurement is closely related to the estimation error.

C. Analysis of Epistemic Uncertainty Estimation

Next, we ignore aleatoric uncertainty and analyze separately epistemic uncertainty in the model parameters. For this, the neural networks are trained to perform only point estimation, i.e., trained with the loss function (8). An ensemble of models is collected by applying Deep ensembles or MC dropout to approximate the predictive mean and variance.

In Fig. 5, we present the same audio example as in Fig. 3 to illustrate the uncertainty measures based on MC dropout and Deep ensembles. MC dropout and Deep ensembles provide the clean speech estimates as shown in the first row of Fig. 5. The estimation error for each method is obtained similarly by calculating the absolute square between the estimated and clean spectral coefficients, shown in the second row. As can be observed, both methods produce large errors as well as associated large uncertainties when the signal is heavily corrupted by noise, i.e., the green box-marked region. While the noise corruption is less severe, i.e., the region marked with a red box, the model generates low prediction errors and also a relatively low level of uncertainty. From the visual analysis, the uncertainty generated by Deep ensembles is more correlated with the error, while MC dropout appears to underestimate the uncertainty of incorrect predictions. To objectively assess the reliability of



(a) Sparsification plot of MC dropout (MC)

(b) Sparsification plot of Deep ensembles (DE)

Fig. 6: Sparsification plots of epistemic uncertainty $\widetilde{\Sigma}$ for the DNS test dataset. The dashed line denotes the lower bound of the corresponding sparsification plot, denoted as Oracle M. A smaller distance of the sparsification plot to the oracle curve indicates a more accurate uncertainty estimation. Note that all oracle curves are visually *overlapping*.

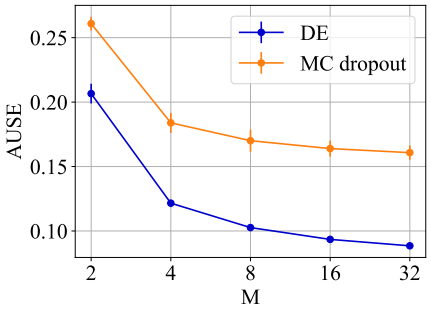


Fig. 7: AUSE for the DNS test dataset. AUSE is plotted relative to a different number of forward passes M. The markers denote the mean and the vertical bars indicate the standard deviation. Lower values indicate a smaller deviation from the oracle curve, and thus more reliable uncertainty estimation.

uncertainty measures, we also utilize the sparsification plots and the sparsification errors, as illustrated in Fig. 6 and Fig. 7 respectively.

In Fig. 6, we show the sparsification plots of Deep ensembles and MC dropout for a different number of forward passes $M \in \{2,4,8,16,32\}$. It can be observed that both MC dropout and Deep ensembles yield decreasing sparsification plots, suggesting that they produce accurate uncertainties that correlate well with the estimation errors. It also shows that a large M leads to a

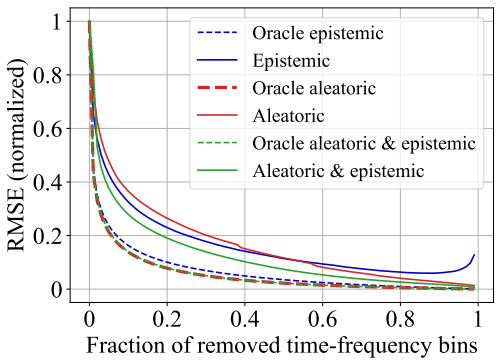


Fig. 8: Sparsification plots of aleatoric $\tilde{\lambda}$, epistemic $\widetilde{\Sigma}$, and overall predictive uncertainty $\widehat{\Sigma}$ (i.e., aleatoric & epistemic) on the DNS test dataset. Note that *Oracle aleatoric* and *Oracle aleatoric & epistemic* overlap.

TABLE I: AUSE values of *Aleatoric*, *Epistemic*, and *Aleatoric & epistemic* in Fig. 8.

	Aleatoric	Epistemic	Aleatoric & epistemic
AUSE	0.110	0.094	0.067

sparsification plot closer to its corresponding oracle curve, i.e., improves the performance of the uncertainty estimation, and this improvement becomes saturated when M is sufficiently large, e.g., from $M=16$ to $M=32$.

To comprehensively compare MC dropout and Deep ensembles in terms of uncertainty modeling, AUSE is plotted as a function of different numbers of forward passes M. Multiple models for each M are used to provide mean and standard deviation to account for variations resulting from random factors in training. 16 MC dropout models are trained and used to compute the mean of AUSE and its standard deviation for each possible M. For Deep ensembles, 16 disjoint sets of M models are randomly selected from the 33 trained models to compute the mean and standard deviation of AUSE. The AUSE plot in Fig. 7 provides an alternative and more informative evaluation than a single sparsification plot. It indicates that Deep ensembles generally produce more accurate uncertainty than MC dropout, which may fail to produce reliable uncertainties for some erroneous predictions. This coincides with our visual observation in the green box-marked region in Fig. 5.

D. Prediction Uncertainty Combining Aleatoric and Epistemic Uncertainties

In this part, we investigate the overall prediction uncertainty obtained by combining aleatoric uncertainty and epistemic uncertainty as in (15). To obtain the overall prediction uncertainty, we use an ensemble of models trained with the optimization metric (12) such that both aleatoric and epistemic uncertainty are captured. It has been shown in Section VI-C that Deep ensembles yield more accurate epistemic uncertainty than MC dropout and, therefore, are selected for the estimation of the overall predictive uncertainty. Although a larger number of models M could potentially improve the mean and variance estimation, we restrict M to 16 as further improvements become subtle while the computation time increases considerably.

In Fig. 8, we use sparsification plots to analyze the quality of prediction uncertainty estimates combining aleatoric and epistemic uncertainties. The corresponding AUSE values are provided in Table I. The plot illustrates that the overall predictive uncertainty estimates correlate stronger with the estimation error than either of the two uncertainties alone. This suggests that two sources of uncertainty may complement each other and combining both leads to more reliable uncertainty estimates. For example, Deep ensembles do not seem to capture sufficient uncertainty for less distorted input (e.g., first three seconds) as shown in Fig. 5, while aleatoric uncertainty shown in Fig. 3 could be able to compensate for this shortcoming.

TABLE II: Evaluation results on the DNS test dataset. All results are stated as mean $\pm$ 95%-confidence interval. *Unc.* stands for *Uncertainty*.

	Unc.	PESQ	ESTOI	SI-SDR
Noisy (DNS)	-	1.58 ± 0.07	0.81 ± 0.02	9.07 ± 0.89
Baseline WF	✗	2.48 ± 0.10	0.90 ± 0.01	16.84 ± 0.74
Baseline SI-SDR	✗	2.63 ± 0.10	0.91 ± 0.01	17.49 ± 0.78
MC dropout	✓	2.53 ± 0.10	0.90 ± 0.01	16.88 ± 0.74
Deep ensembles	✓	2.66 ± 0.10	0.91 ± 0.01	17.16 ± 0.73
Aleatoric-WF	✓	2.62 ± 0.11	0.91 ± 0.01	17.54 ± 0.78
Aleatoric-MAP	✓	2.69 ± 0.10	0.91 ± 0.01	17.54 ± 0.78
DE-Aleatoric-WF	✓	2.77 ± 0.11	0.92 ± 0.01	17.88 ± 0.78
DE-Aleatoric-AMAP	✓	2.83 ± 0.10	0.92 ± 0.01	17.90 ± 0.78

VII. Influence of Modeling Uncertainty for Speech Enhancement Performance

In this section, we show how modeling different sources of uncertainty affects the performance of speech enhancement. To evaluate the speech enhancement performance, we employ perceptual evaluation of speech quality (PESQ) [77] to measure speech quality, extended short-time objective intelligibility (ESTOI) [78] to measure speech intelligibility, and SI-SDR to account for both noise reduction and speech distortion.

To show the impact of modeling aleatoric uncertainty on speech enhancement performance, we compare the performance of the model trained with the proposed loss function (12) with that of Baseline WF and Baseline SI-SDR. The proposed method enables speech estimation via either the Wiener filter, which implicitly takes uncertainty into account during the training process, or the approximated MAP filter, which explicitly includes uncertainty to estimate speech, denoted as Aleatoric-WF and Aleatoric-AMAP respectively. Table II shows the average evaluation results on the DNS synthetic non-reverb test set. Aleatoric-WF shows improvements in PESQ, ESTOI, and SI-SDR compared to the Baseline WF, indicating the benefit of weighting Wiener estimates with uncertainty during training. Further PESQ improvements over both Baseline WF and Baseline SI-SDR can be observed when explicitly incorporating uncertainty into clean speech estimation, that is, Aleatoric-AMAP. This demonstrates the advantage of modeling uncertainty associated with the Wiener estimate rather than directly estimating optimal points. When evaluated on another dataset with speech from WSJ and noise from CHiME3, the performance gap between Aleatoric-AMAP and the baselines in terms of PESQ is further increased, as shown in Fig. 9, indicating that the model that takes uncertainty into account has improved generalization capacities for speech enhancement. This can be attributed to the nonlinear estimation characteristics of the uncertainty-based AMAP estimator with respect to noisy inputs and the resulting better speech preservation properties. We observe larger improvements over the baselines at high SNRs, which might be explained by the

fact that, at high SNRs, speech quality (and thus PESQ) is mainly affected by speech distortions, while at low SNRs the main factor is residual noise. Overall, these evaluation results demonstrate the notable benefits of modeling aleatoric uncertainty in the algorithm.

To show the impact of modeling epistemic uncertainty on speech enhancement performance, we compare the performance of Deep ensembles and MC dropout with Baseline WF. We again restrict M to 16 as in Section VI-D. MC dropout performs comparably to Baseline WF on the DNS test set, while a larger improvement can be observed when using Deep ensembles. This improvement is even more pronounced in PESQ. Similarly, the results on the second test set are shown in Fig. 9, where Deep ensembles and MC dropout improve over Baseline WF in terms of PESQ for all considered SNRs and provide higher ESTOI scores, especially at low SNRs. We observe that Deep ensembles not only provide more accurate uncertainty estimates than MC dropout but also lead to a better speech enhancement performance. A possible explanation is that while MC dropout only captures local uncertainty around a single mode, Deep ensembles trained with different initialization points are capable of exploring multiple modes in the function space to account for training data, see, e.g., [48], [49]. This may allow the neural network to generalize better to complex acoustic scenarios.

To show the impact of modeling predictive uncertainty that combines both aleatoric and epistemic uncertainties on speech enhancement performance, we use the same set of models as described in Section VI-D. We take the average of estimates as in (15) and (16) and obtain two speech estimates, called DE-Aleatoric-WF and DE-Aleatoric-AMAP respectively. They both provide better ESTOI and SI-SDR scores than the baselines, the epistemic uncertainty-only model, and the aleatoric uncertainty-only model, especially at low SNRs. Moreover, DE-Aleatoric-AMAP yields higher scores in PESQ likely due to the uncertainty-dependent regularization and exploration of multiple modes in the function space. This indicates that combining the model that accounts for aleatoric uncertainty with the ensemble-based method can take advantage of the benefits of both approaches and further improve the performance. Overall, the evaluation results across different datasets show that quantifying uncertainty in neural network-based speech enhancement leads to a considerable improvement in enhancement performance over the baseline models.

VIII. Conclusion

In this paper, besides estimating clean speech, we quantified predictive uncertainty in neural network-based speech enhancement. For this, aleatoric uncertainty, which describes inherent uncertainty in data, and epistemic uncertainty, which accounts for uncertainty of the model, were captured and analyzed in a joint framework. We investigated the reliability of uncertainty estimates from different sources, and how it affects the enhancement performance. Our proposed hybrid loss function based on MAP inference of complex spectral coefficients and an AMAP estimator of spectral magnitudes has demonstrated the effectiveness in modeling aleatoric uncertainty. In addition, the proposed scheme provided a principled way to create a noise-removing mask that explicitly incorporates uncertainty to further improve speech enhancement performance. The evaluation results on different datasets have shown increased generalization capacities when modeling aleatoric uncertainty.

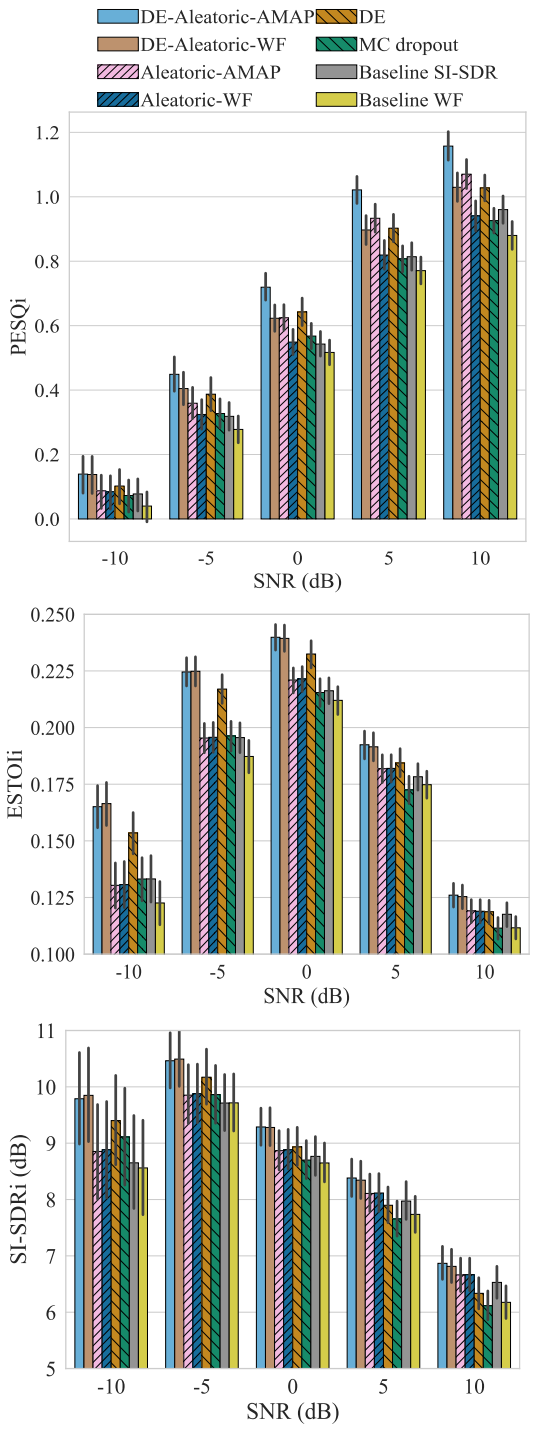


Fig. 9: Performance improvement on the dataset with speech from WSJ0 and noise from CHiME3. PESQi denotes PESQ improvement with respect to noisy mixtures. ESTOIi and SI-SDRi are defined similarly. Markers and vertical bars indicate the mean and 95% confidence interval.

To empirically approximate the predictive distribution and capture epistemic uncertainty, we employed two Bayesian deep learning methods, MC dropout and Deep ensembles. We showed that Deep ensembles not only provide more accurate estimates of epistemic uncertainty than MC dropout, but also lead to more prominent improvements in speech enhancement. A reason may be that Deep ensembles can potentially converge to different local minima in

the loss landscape due to random initialization. Furthermore, we combined the proposed hybrid function with Deep ensembles to quantify overall prediction uncertainty, which reflects both data uncertainty and model uncertainty. An analysis using sparsification plots showed that combining different types of uncertainties further improves the reliability of predictive uncertainty estimation, indicating the complementary nature of the two sources of uncertainty. Finally, our experiments indicated that the performance of clean speech estimation can be considerably improved over the baselines while additionally obtaining predictive uncertainty estimates.

In summary, this work investigated capturing predictive uncertainty in neural network-based speech enhancement and showed the noticeable benefits of modeling uncertainty for clean speech estimation. Uncertainty can indicate the algorithm's confidence in the output in the absence of ground truth, which is essential for assessing the reliability of speech estimates. With this work, we hope to enlighten discussions on modeling uncertainty in the speech enhancement task, while facilitating future research on how to take advantage of uncertainty.

REFERENCES

[1] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[2] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. Wiley, 2018, pp. 65–85.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the ephraim and malah suppression rule for audio signal enhancement," *EURASIP Journal on Advances in Signal Proc.*, vol. 2003, no. 10, pp. 1–9, 2003.

[5] I. Andrianakis and P. White, "MMSE speech spectral amplitude estimators with Chi and Gamma speech priors," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2006, pp. III–III.

[6] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Mar. 2008, pp. 4037–4040.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Proc.*, vol. 9, no. 5, pp. 504–512, 2001.

[8] M. Tammen, D. Fischer, B. T. Meyer, and S. Doclo, "DNN-based speech presence probability estimation for multi-frame single-microphone speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2020, pp. 191–195.

[9] Y. A. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 20, no. 4, pp. 1256–1269, 2012.

[10] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to a priori snr estimation," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 19, no. 1, pp. 186–195, 2011.

[11] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 28, pp. 1404–1415, 2020.

[12] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Apr. 2018, pp. 716–720.

[13] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, "Noisy-target training: A training strategy for dnn-based speech enhancement without clean speech," in *European Signal Proc. Conf. (EUSIPCO)*, 2021, pp. 436–440.

[14] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Int. Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2018, pp. 1–6.

[15] G. Carbajal, J. Richter, and T. Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Jun. 2021, pp. 681–685.

[16] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Jun. 2021, pp. 676–680.

[17] G. Carbajal, J. Richter, and T. Gerkmann, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement," in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, Oct. 2021, pp. 126–130.

[18] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[19] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *Int. Conf. on Telecommunications and Signal Proc. (TSP)*, Jul. 2021, pp. 72–76.

[20] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks," in *Interspeech*, Oct. 2020, pp. 4516–4520.

[21] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2020, pp. 371–375.

[22] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 30, pp. 2993–3007, 2022.

[23] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoders," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 28, pp. 1788–1800, 2020.

[24] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An improved version of MetricGAN for speech enhancement," in *Interspeech*, Aug. 2021, pp. 201–205.

[25] S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, and Y. Tsao, "MetricGAN-U: Unsupervised speech enhancement/dereverberation based only on noisy/reverberated speech," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2022, pp. 7412–7416.

[26] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2022, pp. 7402–7406.

[27] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Interspeech*, Sep. 2022, pp. 2928–2932.

[28] R. Rehr and T. Gerkmann, "SNR-based features and diverse training data for robust DNN-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 29, pp. 1937–1949, 2021.

[29] A. Kumar and D. Florencio, "Speech enhancement in multiple-noise conditions using deep neural networks," in *Interspeech*, Sep. 2016, p. 3738–3742.

[30] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2604–2612, 2016.

[31] 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.

[32] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? Does it matter?" *Structural Safety*, vol. 31, no. 2, pp. 105–112, 2009.

[33] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" *Advances in Neural Information Proc. Systems*, vol. 30, 2017.

[34] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning," in *Int. Conf. Machine Learning (ICML)*, Jul. 2018, pp. 1184–1193.

[35] T. Pearce, A. Brintrup, M. Zaki, and A. Neely, "High-quality prediction intervals for deep learning: A distribution-free, ensembled approach," in *Int. Conf. Machine Learning (ICML)*, Jul. 2018, pp. 4075–4084.

[36] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Proc. Systems*, vol. 30, 2017.

[37] F. K. Gustafsson, M. Danelljan, and T. B. Schon, "Evaluating scalable Bayesian deep learning methods for robust computer vision," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Jun. 2020, pp. 318–319.

[38] E. Ilg, O. Cicek, S. Galesso, A. Klein, O. Makansi, F. Hutter, and T. Brox, "Uncertainty estimates and multi-hypotheses networks for optical flow," in *European Conf. on Computer Vision (ECCV)*, Sep. 2018, pp. 652–667.

[39] L. Chai, J. Du, Q.-F. Liu, and C.-H. Lee, "Using generalized Gaussian distributions to improve regression error modeling for deep learning-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 27, no. 12, pp. 1919–1931, 2019.

[40] K. Kinoshita, M. Delcroix, A. Ogawa, T. Higuchi, and T. Nakatani, "Deep mixture density network for statistical model-based feature enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Mar. 2017, pp. 251–255.

[41] S. M. Siniscalchi, "Vector-to-vector regression via distributional loss for speech enhancement," *IEEE Signal Processing Letters*, vol. 28, pp. 254–258, 2021.

[42] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Int. Conf. Machine Learning (ICML)*, Jun. 2016, pp. 1050–1059.

[43] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *Int. Conf. Machine Learning (ICML)*, Jun. 2011, pp. 681–688.

[44] T. Chen, E. Fox, and C. Guestrin, "Stochastic gradient Hamiltonian Monte Carlo," in *Int. Conf. Machine Learning (ICML)*, Jun. 2014, pp. 1683–1691.

[45] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Int. Conf. Machine Learning (ICML)*, Jun. 2015, pp. 1613–1622.

[46] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," in *Int. Conf. Learning Repr. (ICLR) workshop track*, 2016.

[47] J. Nixon, B. Lakshminarayanan, and D. Tran, "Why are bootstrapped deep ensembles not better?" in *"I Can't Believe It's Not Better!" Neural Information Proc. Systems workshop*, 2020.

[48] S. Fort, H. Hu, and B. Lakshminarayanan, "Deep ensembles: A loss landscape perspective," in *"Bayesian Deep Learning" Neural Information Proc. Systems workshop*, 2019.

[49] A. G. Wilson and P. Izmailov, "Bayesian deep learning and a probabilistic perspective of generalization," *Advances in Neural Information Proc. Systems*, vol. 33, pp. 4697–4708, 2020.

[50] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[51] K. Sridhar and C. Busso, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2020, pp. 8384–8388.

[52] N. Raj Prabhu, G. Carbajal, N. Lehmann-Willenbrock, and G. Timo, "End-to-end label uncertainty modeling for speech-based arousal recognition using Bayesian neural networks," in *Interspeech*, Sep. 2022, pp. 151–155.

[53] S. Braun and S.-C. Liu, "Parameter uncertainty for end-to-end speech recognition," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2019, pp. 5636–5640.

[54] S. Khurana, N. Moritz, T. Hori, and J. L. Roux, "Unsupervised domain adaptation for speech recognition via uncertainty driven self-training," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Jun. 2021, pp. 6553–6557.

[55] A. Vyas, P. Dighe, S. Tong, and H. Bourlard, "Analyzing uncertainties in speech recognition using dropout," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2019, pp. 6730–6734.

[56] K. Alex, B. Vijay, and C. Roberto, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *Proc. of the British Machine Vision Conf. (BMVC)*, Sep 2017, pp. 57.1–57.12.

[57] J. Le Roux, S. Watanabe, and J. R. Hershey, "Ensemble learning for speech enhancement," in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, Oct. 2013, pp. 1–4.

[58] H. Zhang, X. Zhang, and G. Gao, "Multi-target ensemble learning for monaural speech separation," in *Interspeech*, Aug. 2017, pp. 1958–1962.

[59] H. Fang, T. Peer, S. Wermter, and T. Gerkmann, "Integrating statistical uncertainty into neural network-based speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2022, pp. 386–390.

[60] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[61] M. Seitzer, A. Tavakoli, D. Antic, and G. Martius, "On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks," in *Int. Conf. Learning Repr. (ICLR)*, Apr. 2022.

[62] A. Stirn and D. A. Knowles, "Variational variance: Simple, reliable, calibrated heteroscedastic noise variance parameterization," *arXiv:2006.04910*, 2020.

[63] N. Skafte, M. Jørgensen, and S. Hauberg, "Reliable training and estimation of variance networks," *Advances in Neural Information Proc. Systems*, vol. 32, 2019.

[64] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. 28, no. 2, pp. 137–145, 1980.

[65] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2019, pp. 626–630.

[66] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A dissecting approach," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2020, pp. 6359–6363.

[67] R. M. Neal, "Bayesian learning for neural networks," PhD thesis, University of Toronto, 1995.

[68] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," in *Interspeech*, May 2020, pp. 2492–2496.

[69] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Interspeech*, Aug. 2011, pp. 1509–1512.

[70] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Sennheiser LDC93S6B," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[71] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2015, pp. 504–511.

[72] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *ISMIR*, Oct. 2017.

[73] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Interspeech*, Sep. 2018, pp. 3229–3233.

[74] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Jul. 2017, pp. 6924–6932.

[75] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int. Conf. Learning Repr. (ICLR)*, Dec. 2014.

[76] A. S. Wannenwetsch, M. Keuper, and S. Roth, "Probflow: Joint optical flow and uncertainty estimation," in *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Oct. 2017, pp. 1173–1182.

[77] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2001, pp. 749–752.

[78] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 24, no. 11, pp. 2009–2022, 2016.

2.2 Uncertainty Estimation in Deep Speech Enhancement Using Complex Gaussian Mixture Models [P3]

Reference

H. Fang and T. Gerkmann, "Uncertainty estimation in deep speech enhancement using complex Gaussian mixture models," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095213

Copyright Notice

UNCERTAINTY ESTIMATION IN DEEP SPEECH ENHANCEMENT USING COMPLEX GAUSSIAN MIXTURE MODELS

Huajian Fang, Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany

ABSTRACT

Single-channel deep speech enhancement approaches often estimate a single multiplicative mask to extract clean speech without a measure of its accuracy. Instead, in this work, we propose to quantify the uncertainty associated with clean speech estimates in neural network-based speech enhancement. Predictive uncertainty is typically categorized into *aleatoric uncertainty* and *epistemic uncertainty*. The former accounts for the inherent uncertainty in data and the latter corresponds to the model uncertainty. Aiming for robust clean speech estimation and efficient predictive uncertainty quantification, we propose to integrate statistical complex Gaussian mixture models (CGMMs) into a deep speech enhancement framework. More specifically, we model the dependency between input and output stochastically by means of a conditional probability density and train a neural network to map the noisy input to the full posterior distribution of clean speech, modeled as a mixture of multiple complex Gaussian components. Experimental results on different datasets show that the proposed algorithm effectively captures predictive uncertainty and that combining powerful statistical models and deep learning also delivers a superior speech enhancement performance.

Index Terms— Speech enhancement, uncertainty estimation, neural networks, complex Gaussian mixture models

1. INTRODUCTION

Speech enhancement aims to recover clean speech from microphone recordings distorted by interfering noise to improve speech quality and intelligibility. The recordings are often transformed into the time-frequency domain using the short-time Fourier transform (STFT), where an estimator can be applied to extract clean speech. Depending on different probabilistic assumptions, various Bayesian estimators have been presented [1], [2]. A typical example is the Wiener filter derived based on the complex Gaussian distribution of speech and noise signals [2]. Complex Gaussian mixture models (CGMMs) have also been studied in [3] to model super-Gaussian priors, which are considered a better fit for speech signals [2].

Today, deep neural network (DNN)-based approaches are the standard tool for speech enhancement, alleviating shortcomings of traditional methods. Supervised masking approaches are trained on large databases consisting of noisy-clean speech pairs and directly estimate a multiplicative mask to extract clean speech [4]. However, supervised DNN approaches are typically formulated as a problem with a single output, which may result in fundamentally erroneous estimates for unseen samples, without any indication that the erroneous estimate is uncertain. This motivates us to quantify predictive uncertainty associated with clean speech estimates, which allows determining the level of confidence in the outcome in the absence of ground truth.

Predictive uncertainty is typically categorized into *aleatoric uncertainty* and *epistemic uncertainty* [5], [6]. Aleatoric uncertainty describes the uncertainty of an estimate due to the intrinsic randomness of noisy observations. For speech enhancement, it originates from the stochastic nature of both speech and noise. Epistemic uncertainty (also known as *model uncertainty*) corresponds to the uncertainty of the DNN parameters [6]. Hüllermeier et al. [5] provide a general introduction to uncertainty modeling. To quantify aleatoric uncertainty, the dependency between input and output can be modeled stochastically using a speech posterior distribution and enable the DNN to estimate the statistical moments of this distribution. While the predictive mean is a target estimate, the associated variance can be used to measure aleatoric uncertainty [6]. Previous work has implicitly or explicitly explored the uncertainty of aleatoric nature in the context of DNN-based speech enhancement. Chai et al. [7] have proposed a generalized Gaussian distribution to model prediction errors in the log-spectrum domain. Siniscalchi [8] has proposed to use a histogram distribution to approximate the conditional speech distribution, but with a fixed variance assumption, thus failing to capture input-dependent uncertainty. Our previous work [9] allows capturing aleatoric uncertainty based on the complex Gaussian posterior. In contrast, quantifying epistemic uncertainty in the context of DNN-based speech enhancement approaches to account for model's imperfections remains relatively unexplored. In computer vision and deep learning, epistemic uncertainty is usually captured using approximate Bayesian inference. For instance, variational inference can approximate the exact posterior distribution of DNN weights with a tractable distribution [6], [10]. At testing time, multiple sets of DNN weights can be obtained by sampling from an approximate posterior network weight distribution, thus producing multiple different output predictions for each input sample. Epistemic uncertainty captures the extent to which these weights vary given input data, which can be empirically quantified by the variance in these output predictions [6]. However, its computational effort is proportional to the number of sampling passes. This renders those approaches impractical for devices with limited computational resources or strict real-time constraints.

In this work, we propose to integrate statistical CGMMs into a deep speech enhancement framework, so that we can combine the powerful nonlinear modeling capabilities provided by neural networks with super-Gaussian priors as a way to improve the robustness of the algorithm as well as to capture predictive uncertainty. More specifically, we propose to train a DNN to estimate the full posterior distribution of clean speech, modeled as a mixture of multiple complex Gaussian components. The one-to-many mapping based on the CGMM enables the DNN to make multiple reasonable hypotheses, thus increasing the robustness against adverse acoustic scenarios. At the same time, in addition to clean speech estimates, the proposed framework featuring one-to-many mappings allows capturing both aleatoric uncertainty and epistemic uncertainty without extra computational costs. Furthermore, we propose a pre-training scheme to mitigate the mode collapse problem often observed in mixture models, resulting in improved clean speech estimation. Finally, we adapt and employ a gradient modification scheme to effectively stabilize the training of our mixture model.

Note that previous work by Kinoshita et al. [11] also seeks to output multiple hypotheses to avoid deterministic mappings. Our work is different in two main aspects. First, Kinoshita et al. model the logarithm Mel-filterbank features using real Gaussian mixture models, while we follow the prior CGMM of speech and noise spectral coefficients. Second, the DNN outputs in [11] serve as the basis for an additional statistical model-based enhancement method, while we target to obtain clean speech estimates directly via DNNs in an end-to-end fashion.

We thankfully acknowledge the funding from ahoi.digital.

2. SIGNAL MODEL

We consider a single-channel speech enhancement problem in which clean speech is distorted by additive noise. In the STFT domain, the noisy signal is given by

$$X_{ft} = S_{ft} + N_{ft}, \tag{1}$$

where S_{ft} and N_{ft} denote the speech and noise complex coefficients at the frequency bin $f \in \{1,...,F\}$ and the time frame $t \in \{1,...,T\}$. We model the speech and noise signals as mixtures of zero-mean complex Gaussian distributions [3]:

$$S_{ft} \sim \sum_{i=1}^{I} \Omega(i)\mathcal{N}_{\mathbb{C}}(0,\sigma_{i,ft}^2), \quad N_{ft} \sim \sum_{j=1}^{J} \Omega(j)\mathcal{N}_{\mathbb{C}}(0,\sigma_{j,ft}^2). \tag{2}$$

The speech mixture weights $\Omega(i)$ sum to one, and the same applies to the noise mixture weights $\Omega(j)$. The likelihood $p(X_{ft}|S_{ft})$ follows a complex Gaussian mixture distribution centered at S_{ft}, given by

$$p(X_{ft}|S_{ft}) = \sum_{j=1}^{J} \Omega(j) \frac{1}{\pi\sigma_{j,ft}^2} \exp\left(-\frac{|X_{ft}-S_{ft}|^2}{\sigma_{j,ft}^2}\right). \tag{3}$$

Given the speech prior in (2) and the likelihood distribution in (3), one can apply Bayes' theorem to determine the posterior distribution of speech as follows [3]

$$p(S_{ft}|X_{ft}) = \sum_{i=1}^{I}\sum_{j=1}^{J} \Omega(i,j|X_{ft}) \frac{1}{\pi\lambda_{ij,ft}} \exp\left(-\frac{|S_{ft}-W_{ij,ft}^{\text{WF}}X_{ft}|^2}{\lambda_{ij,ft}}\right), \tag{4}$$

where $W_{ij,ft}^{\text{WF}} = \frac{\sigma_{i,ft}^2}{\sigma_{i,ft}^2+\sigma_{j,ft}^2}$ and $\lambda_{ij,ft} = \frac{\sigma_{i,ft}^2\sigma_{j,ft}^2}{\sigma_{i,ft}^2+\sigma_{j,ft}^2}$ are the Wiener filter and the posterior's variance of the mixture Gaussian pair (i,j), respectively. $\Omega(i,j|X_{ft})$ denotes the posterior's mixture weights with the same sum-to-one constraint. The variance $\lambda_{ij,ft}$ for the mixture pair (i,j) can be interpreted as a measure of uncertainty for the Wiener estimate $\widetilde{S}_{ij,ft} = W_{ij,ft}^{\text{WF}}X_{ft}$ [2].

Given an input noisy signal, multiple complex Gaussian components can be combined by computing the expectation of the posterior CGMM, yielding the clean speech estimate

$$\mathbb{E}(S_{ft}|X_{ft}) = \int S_{ft}\, p(S_{ft}|X_{ft}) dS_{ft} = \sum_{i=1}^{I}\sum_{j=1}^{J} \Omega(i,j|X)\widetilde{S}_{ij,ft}\,. \tag{5}$$

The mixture density model possesses the advantage of being able to approximate an arbitrary density function with a sufficient number of components [12], which provides a good fit for modeling, e.g., super-Gaussian characteristics of the speech coefficients. In this work, we propose to embed the CGMM into a DNN framework in order to additionally take advantage of their non-linear modeling capacities, as will be shown next.

3. JOINT ESTIMATION OF CLEAN SPEECH AND PREDICTIVE UNCERTAINTY

Instead of relying on traditional power spectral density tracking algorithms, we can leverage neural networks to directly estimate a mixture of Wiener filters to recover clean speech. Furthermore, it is also possible to optimize the neural network based on the speech posterior distribution (4), so that not only the Wiener filter but also the variance of each mixture pair can be estimated. By taking the negative logarithm and averaging over the time-frequency bins, we can obtain the following optimization problem

$$\widetilde{W}_{l,ft}^{\text{WF}}, \widetilde{\lambda}_{l,ft}, \widetilde{\Omega}_{l,ft} = \underset{W_{l,ft}^{\text{WF}}, \lambda_{l,ft}, \Omega_{l,ft}}{\text{argmin}} \overbrace{-\frac{1}{FT}\sum_{f,t}\log\left(\sum_{l=1}^{L}\exp(\Theta_{l,ft})\right)}^{\mathcal{L}_{p(S|X)}^{\text{CGMM}}},$$

$$\Theta_{l,ft} = \log(\Omega(l|X_{ft})) - \log(\lambda_{l,ft}) - \frac{|S_{ft}-W_{l,ft}^{\text{WF}}X_{ft}|^2}{\lambda_{l,ft}}, \tag{6}$$

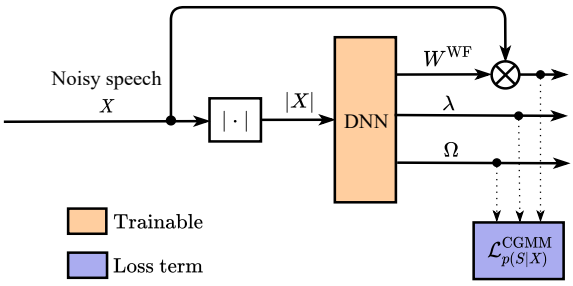


Fig. 1. Block diagram of the DNN-based predictive uncertainty estimation.

where $l \in \{1,...,L\}$ indexes a mixture pair (i,j) in (4), i.e., $L = I \times J$. $\widetilde{W}_{l,ft}^{\text{WF}}$ and $\widetilde{\lambda}_{l,ft}$ denote estimates of the Wiener filter and its associated uncertainty. The CGMM and the corresponding loss function can be viewed as a generalization of the uni-modal Gaussian assumption, which in turn is a generalization of the mean squared error (MSE) loss function. In the limiting case $L = 1$ (i.e., $I = J = 1$), $\mathcal{L}_{p(S|X)}^{\text{CGMM}}$ degenerates into the generic complex Gaussian with a single mean W_{ft}^{WF} and variance λ_{ft}, such that

$$\widetilde{W}_{ft}^{\text{WF}}, \widetilde{\lambda}_{ft} = \underset{W_{ft}^{\text{WF}}, \lambda_{ft}}{\text{argmin}} \underbrace{\frac{1}{FT}\sum_{f,t}\log(\lambda_{ft}) + \frac{|S_{ft}-W_{ft}^{\text{WF}}X_{ft}|^2}{\lambda_{ft}}}_{\mathcal{L}_{p(S|X)}^{\text{CG}}}. \tag{7}$$

Furthermore, by assuming a constant uncertainty for all time-frequency bins and refraining from optimizing for it, $\mathcal{L}_{p(S|X)}^{\text{CG}}$ degenerates into the commonly-used MSE loss [13]:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{FT}\sum_{f,t}|S_{ft}-W_{ft}^{\text{WF}}X_{ft}|^2. \tag{8}$$

In this work, we depart from the uni-modal Gaussian and the constant uncertainty assumption. Alternatively, we propose to map the input to multiple hypotheses by training a DNN with the negative log-posterior $\mathcal{L}_{p(S|X)}^{\text{CGMM}}$, so that we can leverage the better modeling capabilities of the multi-modal distribution and also enable the DNN to quantify the overall predictive uncertainty. Furthermore, incorporating the variances associated with the Wiener estimates enables the adjustment of the weighting of the residual loss, as interpreted in [6], improving the robustness of the network to adverse inputs. As all time-frequency bins are treated independently in (4), the indices ft will be omitted hereafter wherever possible.

We can compute the posterior's variance [14, Section 5], $\text{Var}(S|X)$, to quantify the overall (squared) uncertainty in clean speech estimates originating from different aspects. With the law of total variance, the posterior's variance can be decomposed into [15]:

$$\text{Var}(S|X) = \underbrace{\sum_{l}^{L}\Omega(l|X)\,\lambda_l}_{\mathbb{E}_{l\sim\Omega(l|X)}[\text{Var}(S|X,l)]} + \underbrace{\sum_{l}^{L}\Omega(l|X)\left|W_l^{\text{WF}}X - \mathbb{E}(S|X)\right|^2}_{\text{Var}_{l\sim\Omega(l|X)}(\mathbb{E}[S|X,l])}. \tag{9}$$

The inherent uncertainty associated with the l-th Gaussian component in the outcome is given by $\text{Var}(S|X,l) = \lambda_l$ and aleatoric uncertainty is then quantified as the expectation of variance components $\mathbb{E}_{l\sim\Omega(l|X)}[\text{Var}(S|X,l)]$ following the interpretation in [6], [15]. Epistemic uncertainty can be captured using multiple output predictions, which can be achieved here by the mixture of Wiener estimates, thus circumventing the need for an expensive sampling process. For this, one can compute the variance of the conditional expectation, resulting in the epistemic uncertainty estimate $\text{Var}_{l\sim\Omega(l|X)}(\mathbb{E}[S|X,l])$. Fig 1 depicts an overview of this approach.

Probability density estimation is a non-trivial task. Our preliminary experiments using $\mathcal{L}_{p(S|X)}^{\text{CGMM}}$ directly as the loss function have shown numerical instabilities during training. To overcome this, a gradient modification

scheme inspired by [16] is adapted and employed. Furthermore, DNNs optimized based on $\mathcal{L}^{\text{CGMM}}_{p(S|X)}$ are not guaranteed to exploit the multi-modality of the mixture model, i.e., the multiple hypotheses may converge to the same estimate (collapse to a single mode). We propose to handle this using a pre-training technique based on the winner-takes-all (WTA) scheme [17].

3.1. Gradient modification scheme

The optimization of DNNs with a uni-modal Gaussian (e.g., (7)) as the loss function using stochastic gradient descent shows a high dependence of the gradient on the variance, which is known to cause optimization instabilities [16], [18]. This can be particularly problematic in our CGMM involving multiple complex Gaussian components. It can be seen by computing the gradients of the exponential term Θ_l in (6) with respect to the l-th Wiener filter and associated variance, shown as follows

$$\nabla_{W_l^{\text{WF}}}\Theta_l = \frac{2\text{Re}\{-S\overline{X}+W_l^{\text{WF}}|X|^2\}}{\lambda_l}, \nabla_{\lambda_l}\Theta_l = \frac{\lambda_l-|S-W_l^{\text{WF}}X|^2}{\lambda_l^2}, \quad (10)$$

where the $\text{Re}\{\cdot\}$ operation returns the real part and $\bar{\cdot}$ denotes the complex conjugate. A recent analysis of the real-valued Gaussian assumption by Seitzer et al. [16] showed that the dependence of the gradient on the variance can be reduced by modifying the gradient based on the variance value. Inspired by this, here we extend it to the mixture model, which can be achieved by introducing a weighting term λ^{β_l} to each complex Gaussian component in the loss (6):

$$\widetilde{W}^{\text{WF}}_{l,ft}, \widetilde{\lambda}_{l,ft} = \underset{W^{\text{WF}}_{l,ft},\lambda_{l,ft}}{\text{argmin}} -\frac{1}{FT}\sum_{f,t}\log\left(\sum_{l=1}^{L}\exp(\text{sg}[\lambda^{\beta_l}]\Theta_{l,ft})\right), \quad (11)$$

where $\text{sg}[\cdot]$ denotes the stop gradient operation, which allows λ^{β_l} to act as an input-dependent adaptive factor on the gradient. The parameter $\beta_l \in [0,1]$ controls how much the gradient depends on the l-th variance. As a result, the gradients are modified to

$$\nabla'_{W_l^{\text{WF}}}\Theta_l = \frac{2\text{Re}\{-S\overline{X}+W_l^{\text{WF}}|X|^2\}}{\lambda_l^{1-\beta_l}}, \nabla'_{\lambda_l}\Theta_l = \frac{\lambda_l-|S-W_l^{\text{WF}}X|^2}{\lambda_l^{2-\beta_l}}. \quad (12)$$

Experimentally, we find that the modification in (11) is effective in addressing instability problems during the training of the probabilistic mixture models.

3.2. WTA pre-training scheme

In order to obtain diverse predictions, we propose a pre-training scheme based on the WTA loss [17] to introduce a competition mechanism among the output layers. The concept was originally presented by Guzman-Rivera et al. [17] for support vector machines to produce multiple outputs, and later generalized to the context of DNNs [19]–[22]. We apply the pre-training procedure to a DNN which outputs multiple masks to generate clean speech estimates, i.e., it is equivalent to the CGMM consisting of only the mixture of Wiener estimates. To prompt a network to output diverse hypotheses based on a single ground-truth, the gradient is backpropagated through the top K of the L output predictions at each iteration [21]:

$$\mathcal{L}_{\text{WTA}} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\text{MSE}}(W_k^{\text{WF}}X, S), \quad (13)$$

where the top-K winners are selected based on the MSE measure, indexed by k. Following [21], we start with $K=L$, and gradually halve the number of selections until reaching $K=1$. The competition mechanism prompts the DNN to output diverse clean speech estimates to capture the model's uncertainty, which is expected to alleviate the mode collapse problem to some extent. Previous work has proposed feeding these predictions into a post-processing network to perform distribution fitting [20], [21], while here we propose to use it to initialize the CGMM (except for the output layers that estimates the mixing coefficient and variance of each Gaussian component) to strengthen clean speech estimation without introducing any additional parameters.

Methods	Ale.	Epi.	SNR			Average
			<6 dB	6-12 dB	>12 dB	
Noisy	-	-	1.33/0.74	1.52/0.81	2.00/0.91	1.58/0.81
Baseline WF	✗	✗	2.05/0.85	2.53/0.91	3.03/0.95	2.48/0.90
Prop. CGMM1	✓	✗	2.20/**0.86**	2.66/**0.92**	3.13/**0.96**	2.61/**0.91**
Prop. CGMM4-cons	✗	✓	2.16/**0.86**	2.65/**0.92**	3.17/**0.96**	2.60/**0.91**
Prop. CGMM4	✓	✓	2.21/**0.86**	2.68/**0.92**	3.13/**0.96**	2.62/**0.91**
Prop. CGMM4-pre	✓	✓	**2.22/0.86**	**2.78/0.92**	**3.24/0.96**	**2.69/0.91**

Table 1. Average performance on DNS non-reverb test set. The values are given in PESQ/ESTOI. *Ale.*: Aleatoric; *Epi.*: Epistemic; *Prop.*: Proposed.

4. EXPERIMENTS

4.1. Dataset

We randomly select 80 and 20 hours from the Deep Noise Suppression (DNS) Challenge dataset [23] for training and validation, respectively. The signal-to-noise ratio (SNR) is uniformly sampled between -5 dB and 20 dB. We evaluate the model on two unseen datasets. The first is the non-reverb synthetic test set released by DNS Challenge, which is created by mixing speech signals from [24] with noise from 12 categories [23], at SNRs ranging from 0 dB and 25 dB. We synthesize the second test set by mixing speech samples from WSJ0 (`si_et_05`) [25] and noise samples from CHiME (`cafe`, `street`, `pedestrian`, and `bus`) [26] at SNRs randomly chosen from{-10 dB, -5 dB, 0 dB, 5 dB, 10 dB}.

4.2. Experimental settings

We compute the STFT using a 32 ms Hann window and 50% overlap, at a sampling rate of 16 kHz. For a fair comparison, we base all experiments on a plain U-Net architecture adapted from [27], [28]. The architecture has skip connections between the encoder and the decoder and consists of multiple identical blocks, of which each consists of: 2D convolution layer + instance normalization + Leaky ReLU with slope 0.2. The model processes the inputs of dimension (T,F), with the kernel size $(5,5)$, stride $(1,2)$, and padding $(2,2)$. The encoder is comprised of 6 blocks that increase the feature channel from 1 to 512 progressively ($1-16-32-64-128-256-512$), and then the decoder reduces it back to 16 ($512-256-128-64-32-16-16$), followed by a 1×1 convolution layer that outputs a single mask of the same shape as the input when performing point estimation or outputs L pairs of masks, variance estimates, and mixture weights when applying the CGMM. We set I and J to 2 in (4) , resulting in $L=4$. We set β_l to 0.5 for $l\in\{1,\cdots,L\}$ following [16].

The models are trained using the Adam optimizer with a learning rate of 10^{-3}, which is halved if the validation loss does not decrease for consecutive 3 epochs. Early stopping with a patience of 10 epochs is used. The batch size is 64; the weight decay factor is set to 0.0005. The CGMM can be optionally pre-trained based on the WTA loss as described in Section 3.2. Since it is not straightforward to determine a validation loss for the WTA mechanism, we train the model for 125 epochs with the initial learning rate 10^{-3}, and then halve it every 5 epochs when it is greater than 10^{-6}. We halve the number of winners after every 25 epochs, from $K=4$ to $K=2$, eventually reaching $K=1$, while K remains at 1 for the rest of the training process. The CGMM is then fine-tuned with an initial learning rate of 10^{-5} and the same decay and stopping schemes. Note that the proposed gradient modification scheme described in Section 3.1 is employed to stabilize the training of all CGMM-based networks.

Finally, the following deep algorithms are evaluated:

1. *Baseline WF* refers to a single Wiener filter trained with the loss (8).
2. *CGMM1* refers to the CGMM with $L=1$ (i.e., $I=J=1$) trained using the loss (11). It outputs a single Wiener filter and variance, thus modeling only aleatoric uncertainty.
3. *CGMM4* denotes the CGMM with $L=4$ trained using the loss (11), which captures both aleatoric and epistemic uncertainties.
4. *CGMM4-**cons*** assumes a **cons**tant variance for CGMM4 and refrains from optimizing for it ($\lambda_{l,ft}=1$), capturing epistemic uncertainty

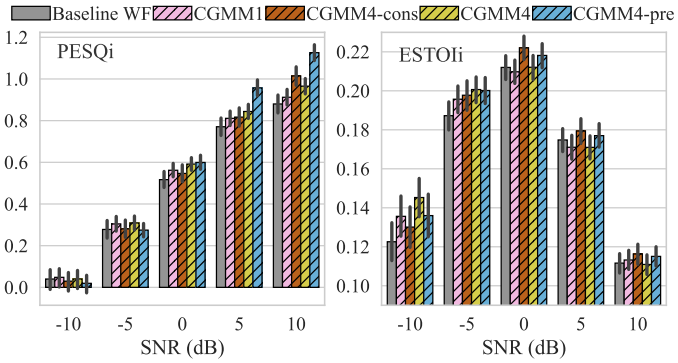


Fig. 2. Performance improvement obtained on the WSJ0-CHiME test set. PESQi denotes PESQ improvement relative to noisy mixtures. The same definition applies to ESTOIi. The marker denotes the mean value and the vertical bar indicates the 95%-confidence interval.

through a mixture of Wiener estimates.

5. *CGMM4-**pre*** refers to the CGMM4 **pre**-trained with the WTA loss.

4.3. Metrics

We present speech enhancement results in terms of perceptual evaluation of speech quality (PESQ) and extended short-time objective intelligibility (ESTOI). We use a sparsification plot [20], [29] to quantitatively evaluate the captured uncertainty. The sparsification plot illustrates the correlation between the uncertainty measures and the true errors. As a first step, the errors of the spectral coefficients are ranked according to their corresponding uncertainty measures. For well-calibrated uncertainty estimates, when the time-frequency bins with large uncertainties are removed the residual error should decrease. Accordingly, the root mean squared error (RMSE) can be plotted versus the fraction of the time-frequency bins removed. Ranking the true errors by their own values yields a lower bound for the sparsification plot, referred to as the *oracle curve*. When the uncertainty estimates and the errors are perfectly correlated, the sparsification plot and the oracle curve overlap.

4.4. Results

In Table 1, we present average evaluation results on the DNS non-reverb test set. It can be observed that the proposed framework considering either aleatoric uncertainty (CGMM1) or epistemic uncertainty (CGMM4-cons) outperforms the point estimation baseline, demonstrating the advantages of modeling uncertainty associated with the clean speech estimates in speech enhancement. Comparing CGMM4 with CGMM1 and CGMM4-cons, the benefits of the modeling both aleatoric and epistemic uncertainties using the multi-modal posterior distribution is not evident. This may be attributed to the fact that training a model based on (11) is not guaranteed to explore the multi-modal modeling capacities of the mixture model. However, this can be largely mitigated by the proposed pre-training scheme, as indicated by the higher PESQ scores of CGMM4-pre.

Fig. 2 shows the improvements of PESQ and ESTOI relative to the noisy mixtures of the synthetic WSJ0-CHiME test set. We observe larger PESQ improvements for the mixture models especially at high input SNRs. In particular, CGMM4-pre yields the highest PESQ improvements, indicating that promoting diverse predictions in the mixture model improves generalization capacities for speech enhancement. Furthermore, it can be observed that the models accounting for uncertainty lead to larger ESTOI improvements at low input SNRs, which again demonstrates the benefits of integrating statistical models into deep speech enhancement as well as modeling uncertainty.

In addition to improving performance, enabling the DNNs to quantify predictive uncertainty is essential to determine how informative a clean speech estimate is without knowing ground-truth (which we do not have access to in practice). Therefore, we evaluate the captured predictive uncertainty in CGMM-pre qualitatively and quantitatively. Fig. 3 shows the spectrograms of an example utterance from the DNS test set. By computing the absolute difference between the clean reference and estimated spectral coefficients, we can measure the prediction error (visualized in the first figure of the second row). It can be observed that both types of uncertainty are closely related to the estimation error, i.e., the model outputs large uncertainties when large errors are produced (e.g., the first 4 seconds of the example utterance). This association is further reflected in Fig. 4, where we observe that both sparsification plots are monotonically decreasing and are close to the oracle curve, implying that both types of uncertainties accurately reflect regions where speech prediction is difficult.

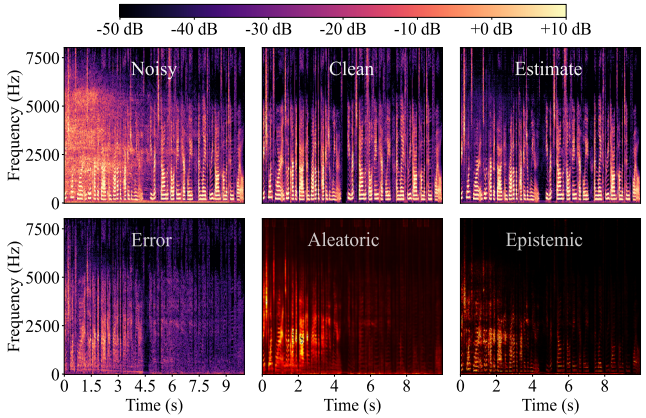


Fig. 3. The uncertainty is visualized as a heatmap, where black indicates low uncertainty and brighter colors indicate higher uncertainty.

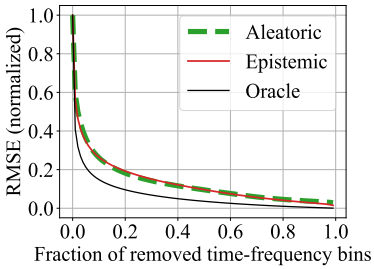


Fig. 4. Sparsification plots created based on all audio samples of the DNS reverb-free synthetic test set. The smaller distance to the oracle curve indicates a more accurate uncertainty estimation.

5. CONCLUSION

In this paper, we have proposed a deep speech enhancement framework to jointly estimate clean speech and quantify predictive uncertainty, based on the statistical CGMM. By estimating the parameters of the full speech posterior distribution involving multiple complex Gaussian components, we can effectively capture both aleatoric and epistemic uncertainties with a single forward pass, circumventing the need for expensive sampling. In addition, the potential of the mixture models can be better exploited if we promote diverse predictions and mitigate the mode collapse problem using the proposed pre-training scheme. Eventually, evaluation results in terms of instrumental measures have demonstrated the considerable advantages of combining powerful statistical models and deep learning compared to directly predicting a point estimate. Our reliable uncertainty estimates can enable interesting future work. For instance, the uncertainty of aleatoric nature can guide multi-modality fusion [30], while epistemic uncertainty capturing the model's ignorance [5] can be used to design a uncertainty-driven training mechanism to improve, e.g., domain adaptation in speech recognition [31].

6. REFERENCES

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement: a survey of the state of the art*. Morgan & Claypool Publishers, 2013, pp. 1–80.

[2] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds., Wiley, 2018, pp. 65–85.

[3] R. F. Astudillo, "An extension of STFT uncertainty propagation for GMM-based super-Gaussian a priori models," *IEEE Signal Proc. Letters*, vol. 20, no. 12, pp. 1163–1166, 2013.

[4] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[5] 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.

[6] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" *Advances in Neural Information Proc. Systems (NIPS)*, vol. 30, 2017.

[7] L. Chai, J. Du, Q.-F. Liu, and C.-H. Lee, "Using generalized Gaussian distributions to improve regression error modeling for deep learning-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 27, no. 12, pp. 1919–1931, 2019.

[8] S. M. Siniscalchi, "Vector-to-vector regression via distributional loss for speech enhancement," *IEEE Signal Proc. Letters*, vol. 28, pp. 254–258, 2021.

[9] H. Fang, T. Peer, S. Wermter, and T. Gerkmann, "Integrating statistical uncertainty into neural network-based speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2022, pp. 386–390.

[10] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," in *Int. Conf. Machine Learning (ICML)*, Jun. 2015, pp. 1613–1622.

[11] K. Kinoshita, M. Delcroix, A. Ogawa, T. Higuchi, and T. Nakatani, "Deep mixture density network for statistical model-based feature enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Mar. 2017, pp. 251–255.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[13] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *Int. Conf. on Telecommunications and Signal Processing (TSP)*, Jul. 2021, pp. 72–76.

[14] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4.

[15] S. Choi, K. Lee, S. Lim, and S. Oh, "Uncertainty-aware learning from demonstration using mixture density networks with sampling-free variance modeling," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, May 2018, pp. 6915–6922.

[16] M. Seitzer, A. Tavakoli, D. Antic, and G. Martius, "On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks," in *Int. Conf. Learning Repr. (ICLR)*, Apr. 2022.

[17] A. Guzman-Rivera, D. Batra, and P. Kohli, "Multiple choice learning: Learning to produce multiple structured outputs," *Advances in Neural Information Proc. Systems (NIPS)*, vol. 25, 2012.

[18] H. Takahashi, T. Iwata, Y. Yamanaka, M. Yamada, and S. Yagi, "Student-t variational autoencoder for robust density estimation," in *Int. Joint Conf. on Artificial Intelligence (IJCAI)*, Jul. 2018, pp. 2696–2702.

[19] S. Lee, S. Purushwalkam Shiva Prakash, M. Cogswell, V. Ranjan, D. Crandall, and D. Batra, "Stochastic multiple choice learning for training diverse deep ensembles," *Advances in Neural Information Proc. Systems (NIPS)*, vol. 29, pp. 2127–2135, 2016.

[20] E. Ilg, O. Cicek, S. Galesso, *et al.*, "Uncertainty estimates and multi-hypotheses networks for optical flow," in *European Conf. on Computer Vision (ECCV)*, Jan. 2018, pp. 652–667.

[21] O. Makansi, E. Ilg, O. Cicek, and T. Brox, "Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Jun. 2019, pp. 7144–7153.

[22] K. Panousis, S. Chatzis, and S. Theodoridis, "Nonparametric Bayesian deep networks with local competition," in *Int. Conf. Machine Learning (ICML)*, May 2019, pp. 4980–4988.

[23] C. K. Reddy, V. Gopal, R. Cutler, *et al.*, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Interspeech*, Oct. 2020, pp. 2492–2496.

[24] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Twelfth Annual Conf. of the Int. Speech Communication Association*, Aug. 2011, pp. 1509–1512.

[25] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Sennheiser LDC93S6B," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[26] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2015, pp. 504–511.

[27] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, Oct. 2017.

[28] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Interspeech*, Sep. 2018, pp. 3229–3233.

[29] A. S. Wannenwetsch, M. Keuper, and S. Roth, "Probflow: Joint optical flow and uncertainty estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Oct. 2017, pp. 1173–1182.

[30] 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," *arXiv preprint arXiv:2206.05833*, 2022.

[31] S. Khurana, N. Moritz, T. Hori, and J. Le Roux, "Unsupervised domain adaptation for speech recognition via uncertainty driven self-training," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Jun. 2021, pp. 6553–6557.

2.3 Uncertainty-Based Remixing for Unsupervised Domain Adaptation in Deep Speech Enhancement [P4]

Reference

H. Fang and T. Gerkmann, "Uncertainty-based remixing for unsupervised domain adaptation in deep speech enhancement," in *18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aalborg, Denmark, 2024, pp. 45-49. DOI: 10.1109/IWAENC61483.2024.10694221

Copyright Notice

Term Clarification:

In [P4], we refer to spectral masking-based algorithms, which learn a DNN to predict filter masks from inputs using a labeled dataset, as discriminative approaches. The term discriminative modeling is primarily associated with classification tasks, although in a broader sense, this concept can also be applied to regression tasks. In this thesis, we use a unified term, predictive modeling, to encompass both the regression and classification settings, as discussed in [11].

UNCERTAINTY-BASED REMIXING FOR UNSUPERVISED DOMAIN ADAPTATION IN DEEP SPEECH ENHANCEMENT

Huajian Fang, Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany

ABSTRACT

Recent work has shown the effectiveness of remixing-based unsupervised domain adaption algorithms, where a student model is fine-tuned on self-labeled noisy-clean speech data synthesized by remixing speech and noise predictions from the teacher model. However, the optimization of the student model may be hindered by learning from fundamentally erroneous pseudo-targets created by the teacher model. To address this limitation, we augment the teacher model with an uncertainty estimation task and propose an uncertainty-based remixing method that allows the student model to learn from the teacher model's high-quality speech estimates and effectively suppress noise. Experiments demonstrate improved robustness against data mismatches between training and testing conditions, especially for challenging inputs with low signal-to-noise ratios. Moreover, by adjusting the uncertainty threshold to categorize the teacher's estimates for unlabeled noisy samples as reliable or unreliable, the proposed uncertainty-based remixing process allows for a controllable trade-off between noise suppression and speech preservation, enabling the model to be adapted to diverse application needs.

Index Terms— Speech enhancement, deep learning, domain adaptation, uncertainty modeling

1. INTRODUCTION

Speech recorded by microphones is inevitably distorted by ambient noise, which causes difficulties for digital communication devices to work reliably in noisy environments. This requires a speech enhancement system that extracts the target speech from the corresponding noisy mixture. While traditional methods exhibit limited performance for challenging acoustic inputs [1], in recent years, the field has seen great strides through deep learning-based algorithms [2], which can leverage powerful non-linear modeling capacities of neural networks.

Discriminative supervised learning has emerged as a dominant technique in speech enhancement, where neural networks are trained to learn a mapping relationship between noisy mixtures and clean speech using a labeled training dataset consisting of large amounts of noisy-clean speech pairs. However, data mismatch between the source domain (in which the model is trained) and the target domain (to which the model is applied) may raise generalization issues, that is, the performance of speech enhancement algorithms may degrade significantly when tested on noisy samples that do not match the training data. When the domain mismatch is severe, the performance degradation can be particularly large. While the performance of algorithms can potentially be improved by fine-tuning the network model with newly collected paired data in the target domain, this pairwise data collection process is often cumbersome and may require careful and costly post-processing after recording. In contrast, collecting unlabeled noisy mixtures is more feasible, and given enough time, the quantity of noisy data can be virtually infinitely large. In this work, we investigate how to alleviate performance degradation caused by domain mismatch without accessing the target domain's ground truth (i.e., clean speech), which is referred to as *unsupervised target domain adaptation*.

Improving the robustness of algorithms using unlabeled data from the target domain has been an active research topic in deep learning [3]–[6]. Existing methods have successfully developed various effective image [4]–[6] and speech [7]–[9] signal processing models utilizing unlabeled data, mainly for classification tasks. Recent work has demonstrated a growing interest in how unlabeled data can be utilized in complex regression tasks, such as in source separation and speech enhancement [10]–[14]. Sivaraman et al. create a pseudo-paired dataset by mixing noisy recordings with isolated noise signals and optimize network models with a loss function designed to down-weight the contribution of noisy ground truth according to estimated input signal-to-noise ratios (SNRs) [15]. Wisdom et al. [16] propose mixture invariant training (MixIT) for universal sound separation, which forms the input to the separation model by summing easy-to-collect acoustic mixtures. The unsupervised training relies on the independence assumption between sources to perform sound separation. Other work has also explored similar non-parallel training settings for speech enhancement, e.g., [17], [18]. While training the model directly with noisy mixtures of the target domain may circumvent performance degradation between training and testing, achieving performance on par with conventional supervised algorithms remains challenging.

A more advantageous and practical setting is to leverage paired data from the source domain for initial learning, followed by adapting the model to the target domain through unsupervised methods, allowing combining benefits of supervised and unsupervised training [10]–[13], [19], [20]. For example, a teacher-student learning framework has been applied to unsupervised personalized speech enhancement [10], where a large teacher model is pre-trained on a labeled dataset and then generates pseudo-clean targets from unlabeled mixtures in the target domain to train a specialized compact student model. In contrast, Wang et al. [11] use the same architecture for both teacher and student models in singing voice separation. The source estimates of the pre-trained teacher model are randomly remixed to generate self-labeled mixtures offline to train the student model. Similar remixing-based teacher-student frameworks have also been proposed in speech enhancement to leverage unlabeled data to alleviate generalization issues, e.g., [12], [13]. Tzinis et al. [13] improve the diversity of artificial mixtures by online remixing teacher's speech and noise estimates in a single batch and train the student model similarly by treating teacher's estimates as the ground-truth labels. Lam et al. [12] propose to generate self-labeled mixtures to train the student model by remixing the teacher's source estimates at random SNRs. However, when the teacher model's estimates are unreliable and erroneous, the student model is forced to match fundamentally incorrect pseudo-targets, thus negatively affecting the learning process of the student model.

In this paper, we follow the idea of remixing-based domain adaptation [13] and present an uncertainty-based data augmentation process to prevent the student model from learning from erroneous pseudo-labels generated by the teacher model. For this, we incorporate uncertainty estimation into the teacher model, such that the neural network can provide not only clean speech estimates but also the associated confidence (or *uncertainty*) [21], [22]. Uncertainty indicates discrepancies

between predictions and the true data [23] and thus can be used to assess the quality of the estimates. Specifically, the teacher model pre-trained on the source-domain data first generates speech and noise estimates for unlabeled noisy samples in the target domain. In the next step, we filter out low-quality (*unreliable*) speech estimates based on uncertainty estimates and remix only the selected high-quality (*reliable*) speech estimates with noise estimates. Our experiments have demonstrated that the uncertainty-based remixing process exhibits improved robustness against data mismatch between training and testing conditions. Interestingly, we show that by adjusting the uncertainty threshold to distinguish between reliable and unreliable estimates, the proposed uncertainty-based remixing method can achieve different trade-offs between noise reduction and speech distortion, and thus can be flexibly adapted to application scenarios with different requirements.

2. SUPERVISED SPEECH ENHANCEMENT

In this work, we consider the speech enhancement problem in the time-frequency domain and use the short-time Fourier transform (STFT) to convert time-domain signals into their time-frequency representations. It is generally assumed that clean speech recorded by a single microphone is distorted by additive noise, resulting in a speech-plus-noise model:

$$X_{ft} = S_{ft} + N_{ft}. \tag{1}$$

X_{ft}, S_{ft}, and N_{ft} are complex spectral coefficients of the noisy mixture, clean speech, and noise, respectively. The frequency bin and time frame indices are given by $f \in \{1,2,\cdots,F\}$ and $t \in \{1,2,\cdots,T\}$, respectively. Clean speech estimates can be typically obtained by applying a multiplication mask, denoted by W_{ft}, to time-frequency representations of noisy mixtures, followed by the inverse STFT.

Bayesian modeling considers spectral coefficients of speech and noise as realizations of random variables and provides a series of principled methods to derive filter masks. Assuming that speech and noise are independent and follow a circularly symmetric complex Gaussian distribution: $S_{ft} \sim \mathcal{N}_{\mathbb{C}}(0,\sigma^2_{S,ft})$ and $N_{ft} \sim \mathcal{N}_{\mathbb{C}}(0,\sigma^2_{N,ft})$, we can obtain the posterior distribution of speech spectral coefficients:

$$p(S_{ft}|X_{ft}) = \frac{1}{\pi\lambda_{ft}} \exp\left(-\frac{|S_{ft} - W_{ft}X_{ft}|^2}{\lambda_{ft}}\right), \tag{2}$$

where $W^{\text{WF}}_{ft} = \frac{\sigma^2_{s,ft}}{\sigma^2_{s,ft}+\sigma^2_{n,ft}}$ is referred to as Wiener filter and $\lambda_{ft} = \frac{\sigma^2_{s,ft}\sigma^2_{n,ft}}{\sigma^2_{s,ft}+\sigma^2_{n,ft}}$ is the variance of the posterior distribution. While traditional algorithms can use power spectral density tracking algorithms to estimate the variances of speech and noise to implement Wiener filtering, more recent speech enhancement work leverages the non-linear modeling capabilities of neural networks to estimate multiplicative filter masks, W_{ft}, directly. This is achieved by training a neural network using a labeled dataset consisting of large amounts of paired noisy-clean speech data [2]. The neural network is optimized to learn a mapping relationship between the noisy mixtures and the corresponding clean speech, referred to as *supervised discriminative approaches*.

2.1. Deep Uncertainty Estimation

The performance of supervised discriminative methods is often related to the diversity and quantity of training data. Since the synthetic training dataset cannot fully replicate the acoustic conditions of the target domain, there is an inevitable mismatch between training and testing data. This data discrepancy often results in performance degradation, especially for noisy samples with acoustic conditions largely deviated from training data. As a result, deep discriminative methods that predict a single filter mask may generate fundamentally erroneous estimates for unseen samples without any indication of error. To address this issue, we can quantify predictive uncertainty associated with clean speech estimates, which allows us to assess the confidence of the outcome without access to the true data [21], [22], [24].

To augment the network model with an uncertainty estimation task, we can optimize the parameters of the neural network by minimizing the negative log-speech posterior:

$$\widetilde{W}_{ft}, \widetilde{\lambda}_{ft} = \underset{W_{ft},\lambda_{ft}}{\operatorname{argmin}} \underbrace{\frac{1}{FT}\sum_{f,t} \log(\lambda_{ft}) + \frac{|S_{ft} - W_{ft}X_{ft}|^2}{\lambda_{ft}}}_{\mathcal{L}_{p(S|X)}}, \tag{3}$$

where $\widetilde{W}_{ft}$ and $\widetilde{\lambda}_{ft}$ denote estimates of the Wiener filter and the associated variance. The variance λ_{ft} can be interpreted as a measure of uncertainty associated with the minimum mean squared error (MMSE) speech estimator [25], [26]. When uncertainty for each time-frequency bin is assumed to be constant and is not optimized, the negative log-speech posterior loss $\mathcal{L}_{p(S|X)}$ degenerates into the widely-used mean squared error (MSE) loss:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{FT}\sum_{f,t} |S_{ft} - W_{ft}X_{ft}|^2 \tag{4}$$

In this work, we investigate the challenging topic of fine-tuning a pre-trained network model using only the target-domain noisy mixtures. For this, we follow the remixing-based teacher-student learning [12], [13], where the teacher model is pre-trained on a source-domain dataset consisting of parallel noisy-clean speech data. Different from previous work, here we augment the teacher model to output speech estimates and associated uncertainty estimates by training the neural network with the loss (3). This is achieved by splitting the neural network output layer into two, for the mask W_{ft} and uncertainty λ_{ft} respectively. More details on uncertainty estimation in deep speech enhancement can be found in, e.g., [21], [22]. The student model uses the same architecture as the teacher model, except that the student model has only one output layer for filter mask estimation. In the following section, we will present how the student model can learn from noisy mixtures by performing online remixing of teacher's speech and noise estimates [13], and furthermore, how to incorporate uncertainty estimates from the teacher model to prevent the student model from learning from erroneous pseudo-targets.

3. UNSUPERVISED DOMAIN ADAPTATION

To exploit unlabeled noisy mixtures in the target domain, we follow the teacher-student training method proposed in [13]. In this method, the teacher model is pre-trained on a labeled source-domain dataset in a conventional supervised manner. To adapt to the unlabeled target domain, the pre-trained teacher model first generates speech and noise estimates from a batch of unlabeled noisy mixtures, shown as $\widehat{S}_b, \widehat{N}_b = \text{TeacherModel}(X_b)$, where $X_b \in \mathbb{R}^{B\times F\times T}$, $\widehat{S}_b \in \mathbb{R}^{B\times F\times T}$, and $\widehat{N}_b \in \mathbb{R}^{B\times F\times T}$ denote a batch of noisy mixtures, speech estimates, and noise estimates respectively. B denotes the batch size. Then, we randomly permute (RP) the noise estimates and add them to clean speech estimates to generate a bootstrapped batch of noisy mixtures, shown as $\widehat{X}_b = \widehat{S}_b + \text{RP}(\widehat{N}_b)$. The new batch of noisy mixtures is finally used to train the student model: $\widetilde{S}_b, \widetilde{N}_b = \text{StudentModel}(\widehat{X}_b)$. As the bootstrapped batch of noisy mixtures is self-labeled by the teacher model, the student model is optimized by treating the teacher's estimates as ground truth, that is, the loss function (e.g., MSE (4)), is computed between $\widehat{S}_b$ and $\widetilde{S}_b$. Moreover, the teacher model can be updated during the training process using, e.g., an exponential moving average update rule [27], thus making the teacher-student framework a continuous learning process. This method is referred to as *blind remixing* in Section 4.3.

However, due to the data shift between the source and target domains, the teacher model may perform poorly for unseen inputs with different noise types and SNRs, etc [28], [29]. Consequently, this can lead the student to match fundamentally erroneous estimates generated by the teacher model, which is particularly problematic when the data mismatch is large. In the extreme case, when the teacher model generates clean speech estimates with severe noise leakage and the noise estimates suffer similarly from speech leakage, the student model trained on the bootstrapped batches can be misled into learning to extract noise and suppress speech. In this work, we attempt to address this issue by incorporating uncertainty into the teacher model, where we filter out the low-quality speech estimates and enable the student model to learn from high-quality pseudo-targets, as will be detailed next.

3.1. Uncertainty-Based Domain Adaptation

The teacher model is trained using the speech log-posterior loss function (3). As described in Section 2.1, the neural network can simultaneously estimate a multiplicative mask and the associated uncertainty in a manner grounded in Bayesian modeling. This stands in contrast to heuristic quality metrics, such as voice activity-based [2] and segmental SNR-based [15] methods, which require an extra neural network to perform pre-defined auxiliary tasks to approximate the quality assessment of estimates. Uncertainty-based domain adaption can be performed in two steps.

Uncertainty-based data selection step. For each noisy sample in an unlabeled target-domain dataset, the teacher model can provide an utterance-level uncertainty value for the entire estimated clean speech estimate. This is achieved by averaging the associated uncertainty values across the time-frequency bins of the sample. Due to the additive modeling assumption, the noise estimate can be obtained by subtracting the estimated clean speech from the noisy mixture in the time domain. Consequently, lower uncertainty estimates indicate higher quality in both speech and noise estimates, whereas higher uncertainty estimates suggest lower quality. After processing all unlabeled noisy mixtures from the target-domain dataset using the pre-trained teacher model, the noisy mixtures can be sorted based on the uncertainty estimates assigned by the teacher model to the corresponding speech and noise estimates. Given a sorted array of noisy mixtures, we can categorize the noisy mixtures with uncertainty estimates less than a predefined threshold into the high-quality (reliable) group and the rest into the low-quality (unreliable) group. Correspondingly, the resulting speech and noise estimates obtained by the teacher model can be expressed as reliable speech (RS), reliable noise (RN), unreliable speech (US), and unreliable noise (UN) estimates. The student model trained in blind bootstrapped remixing [13] may be adversely affected by low-quality remixed noisy mixtures, such as US-plus-RN and US-plus-UN. As discussed previously, US-plus-UN may mislead the student model into learning to extract noise and suppress speech, while similarly, US-plus-RN can cause the student model to learn to extract noise from noise.

Uncertainty-based data remixing step. The uncertainty-based remixing strategy aims to prevent the student model from learning from low-quality speech estimates. This is achieved by filtering out US estimates and remixing reliable speech estimates with all noise estimates, giving RS-plus-RN and RS-plus-UN. Clearly, remixing RS with RN can produce high-quality bootstrapped noisy mixtures. In contrast, RS-plus-UN may seem to be less optimal. However, the high-uncertainty speech and noise estimates may occur on noisy samples with challenging input SNRs, i.e., noise-dominated mixtures. Thus, the UN estimates may contain substantial noise distortion that may be useful when remixed with reliable pseudo-clean targets, i.e., RS. Overall, it remains crucial to incorporate a broad range of noise estimates since more noise estimates can provide rich acoustic characteristics of the target domain, thereby improving its noise reduction capabilities.

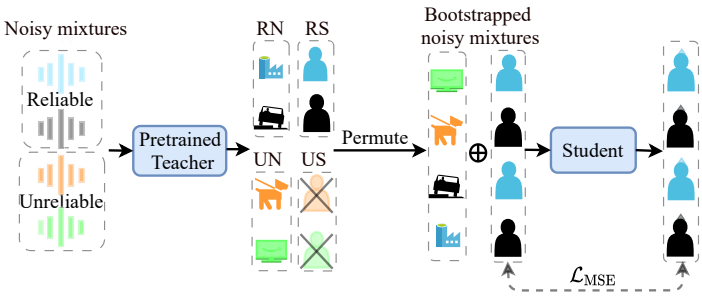


Fig. 1. Illustration of uncertainty-based remixing process. *RN* and *RS* stand for reliable noise and reliable speech estimates. *UN* and *US* represent unreliable noise and unreliable speech estimates.

Given a dataset where the noisy mixtures are sorted according to the corresponding uncertainty estimates, the uncertainty-based remixing step can be achieved by two batch samplers with the same batch size, as illustrated in Figure 1. We define a threshold to classify noisy mixtures associated with low-uncertainty estimates into the reliable group, from which the first batch sampler yields noisy samples. The second batch sampler samples from the rest of the noisy files, i.e., the unreliable group. The speech and noise estimates from the first batch sampler are used to produce a bootstrapped batch of noisy mixtures, i.e., RS-plus-RN, whereas the speech estimates from the second batch, US, are abandoned, while the corresponding noise estimates, UN, are kept. To exploit the noise characteristics contained in UN, we repeat reliable speech estimates in the same step and remix them with unreliable noise estimates, resulting in RS-plus-UN. With this, the uncertainty-based remixing strategy aims to learn from high-quality pseudo-targets generated by the teacher model, while at the same time learning to discriminate noise characteristics of the target domain.

4. EXPERIMENTS

4.1. Datasets

We use the Deep Noise Suppression (DNS) Challenge database [30] as the source-domain data and randomly select a subset consisting of 45 hours of noisy mixtures. These noisy mixtures are created with SNRs uniformly sampled from{-5 dB, -4 dB, ···, 15 dB}. We create a target-domain dataset whose acoustic conditions differ from the source domain, including both noise type and the range of SNR. We mix clean speech from the `si_tr_s` subset of the WSJ0 dataset [31] with noise clips from CHiME3 [32] at SNRs sampled from{-8 dB, -7 dB, ···, 8 dB}. Instead of uniform sampling, the SNR values of the target-domain data follow a truncated normal distribution with a mean of -5 dB and a standard deviation of 5 dB. The amount of target-domain noisy mixtures is the same as the source domain, i.e., 45 hours. The loudness levels of synthetic noisy mixtures are re-scaled to range from -35 dBFS to -15 dBFS. Note that when fine-tuning the student model with the target-domain data, only the noisy mixtures are accessible. For evaluation of unsupervised domain adaptation, we create a test set using clean speech signals from `si_et_05` subset of the WSJ0 and four types of noise signals from CHiME3 (cafe, street, pedestrian, and bus) with SNRs randomly sampled from{-10 dB, -5 dB, 0 dB, 5 dB, 10dB}. Note that there is no data overlap between training and testing.

4.2. Settings

We convert the time-domain signals into their time-frequency representations using the STFT with a 32 ms Hann Window and 50 % overlap.

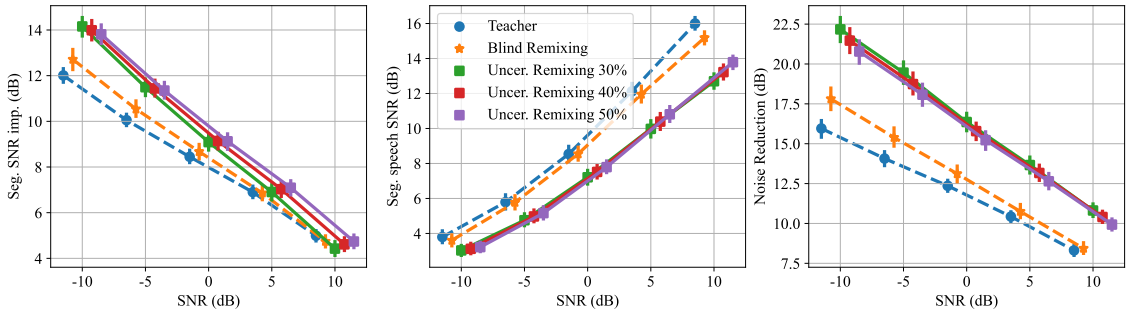


Fig. 2. Performance evaluated at different input SNRs. *Seg.* and *Uncer.* stand for *Segmental* and *Uncertainty*. Markers and error bars denote the mean values and 95%-confidence intervals. For all measures, higher values indicate better performance.

All speech and noise signals are sampled at 16 kHz.

We evaluate the unsupervised domain adaptation using a lightweight U-Net architecture adapted from [33], [34]. The U-Net architecture features an encoder and a decoder, each composed of 5 blocks. The model processes the magnitude spectrogram of dimension (T, F). The encoder/decoder blocks consist of 2D convolution/transposed convolution layers with the kernel size (2, 3) and stride (1, 2), followed by leaky ReLU activation ($\alpha = 0.2$). The encoder increases the feature channel progressively (8-16-32-64-64) and the decoder decreases it back to 8 (64-32-16-8-8), followed by a convolution layer of the kernel (1, 3) to output a filter mask of the same shape as the input. For the teacher model that performs uncertainty estimation, we add an extra convolution layer with the same parameters as the mask branch to output the variance estimates. Moreover, the skip connections between the encoder and decoder are achieved by 1x1 convolutions and addition [34]. The model architecture is implemented using causal convolutions with approximately 88K parameters. Since the neural network is based on causal operations, no normalization techniques using global statistics are applied to the noisy inputs.

The teacher model is pre-trained on the source-domain data using the Adam optimizer with a learning rate of 0.0005 for 150 epochs. Since optimizing a teacher model with the probabilistic loss (3) may lead to numerical instabilities, we use the gradient adaption scheme as discussed in [22]. The model performing best on the validation set is saved. The student model is initialized with the weights of the pre-trained teacher model (except the uncertainty estimation layer, as the student model has only a mask estimation layer focusing only on speech enhancement) and is fined-tuned using a learning rate of 0.0001 for 60 epochs. The learning rate is reduced to one-third of its current value every 15 epochs. During fine-tuning, the teacher model is updated using the exponential moving average update technique with a momentum value of 0.99 [13]. The two batch samplers have a batch size of 32, leading to a total batch size of 64. All adaptation models are trained for the same number of steps per epoch, independent of the threshold used to classify reliable/unreliable groups.

The proposed uncertainty-based remixing is compared with two baselines: 1) the supervised teacher model trained exclusively on the source-domain data, and 2) the domain adaptation framework based on blind remixing [13]. We evaluate the performance of the methods using segmental speech SNR, noise reduction, and segmental SNR improvement as outlined in [1]. They are measures for speech distortion, noise reduction, and a combined assessment of speech distortion and noise reduction, respectively.

4.3. Experimental Results

The results of evaluations are presented in Figure 2. For the proposed uncertainty-based remixing method, we rank the noisy mixtures from lowest to highest based on uncertainty estimates assigned by the teacher model to their corresponding speech and noise estimates, as described in Section 3.1. We then categorize the first 30%, 40%, or 50% of the noisy mixtures into the reliable group (we omit reporting other possible thresholds to prevent clutter). We can observe that the uncertainty-based remixing strategy results in a better trade-off between speech distortion and noise reduction than the blind remixing baseline [13]. The uncertainty-based remixing method leads to better noise reduction performance at the cost of speech distortion, but also higher segmental SNR improvement, especially at low input SNRs. Furthermore, increasing the threshold to incorporate more noisy mixtures into the reliable group (e.g., from 30% to 50%) leads to the uncertainty-based remixing preserving more speech at the expense of noise reduction, as shown by lower noise reduction scores and higher segmental speech SNR scores. This may be because incorporating a larger percentage of noisy mixtures can increase the likelihood of using erroneous speech estimates as pseudo-clean targets, thus converging to the blind remixing baseline. This observation is particularly interesting because it allows the methods to be tailored to different application scenarios and specific noise reduction requirements.

In addition, we also evaluate the performance of the setting where only RS-plus-RN is adopted during the remixing step. We observe that the overall improvement over the blind remixing baseline is marginal (thus, the results are not reported). We hypothesize that while the quality of the pseudo mixtures is important, the quantity of data also plays an important role in the unsupervised domain adaptation. This also demonstrates the benefits of the proposed strategy which reuses a broad spectrum of noise estimates from the unreliable group. Future work may evaluate the scheme with larger and more diverse datasets.

5. CONCLUSION

While supervised speech enhancement has demonstrated good performance, ensuring robustness to unseen acoustic conditions during testing remains challenging. In this work, we explored the problem of performance degradation in masking-based supervised speech enhancement due to data mismatch between training and testing. We presented a method to incorporate uncertainty modeling into remixing-based unsupervised domain adaptation. By filtering out low-quality pseudo-targets generated by the teacher model, the student model learns only from high-quality speech estimates. The uncertainty-based remixing also allows the student model to effectively learn representative noise characteristics of the target domain by repeating reliable speech estimates in the same step and remixing them with low-quality noise estimates, leading to better noise reduction capability at the cost of speech distortion. In addition, different trade-offs between noise reduction and speech distortion can be observed by adjusting the uncertainty threshold.

6. REFERENCES

[1] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 20, no. 4, pp. 1383–1393, 2011.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] V. Verma, K. Kawaguchi, A. Lamb, *et al.*, "Interpolation consistency training for semi-supervised learning," *Neural Networks*, vol. 145, pp. 90–106, 2022.

[4] K. Sohn, D. Berthelot, N. Carlini, *et al.*, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[5] D. Berthelot, R. Roelofs, K. Sohn, N. Carlini, and A. Kurakin, "Adamatch: A unified approach to semi-supervised learning and domain adaptation," in *Int. Conf. Learning Repr. (ICLR)*, 2022.

[6] D. Berthelot, N. Carlini, E. D. Cubuk, *et al.*, "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *Int. Conf. Learning Repr. (ICLR)*, 2020.

[7] D. S. Park, Y. Zhang, Y. Jia, *et al.*, "Improved Noisy Student Training for Automatic Speech Recognition," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 2817–2821.

[8] J. Li, M. L. Seltzer, X. Wang, R. Zhao, and Y. Gong, "Large-Scale Domain Adaptation via Teacher-Student Learning," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 2386–2390.

[9] S. Khurana, N. Moritz, T. Hori, and J. Le Roux, "Unsupervised domain adaptation for speech recognition via uncertainty driven self-training," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 6553–6557.

[10] S. Kim and M. Kim, "Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation," in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, New Paltz, NY, USA, 2021, pp. 176–180.

[11] Z. Wang, R. Giri, U. Isik, J.-M. Valin, and A. Krishnaswamy, "Semi-supervised singing voice separation with noisy self-training," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 31–35.

[12] M. W. Lam, J. Wang, D. Su, and D. Yu, "Mixup-breakdown: A consistency training method for improving generalization of speech separation models," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 6374–6378.

[13] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, and A. Kumar, "Continual self-training with bootstrapped remixing for speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Singapore, Singapore, 2022, pp. 6947–6951.

[14] J. Han and Y. Long, "Heterogeneous separation consistency training for adaptation of unsupervised speech separation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, p. 6, 2023.

[15] A. Sivaraman, S. Kim, and M. Kim, "Personalized Speech Enhancement Through Self-Supervised Data Augmentation and Purification," in *Proc. Interspeech*, Brno, Czech Republic, 2021, pp. 2676–2680.

[16] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," *Advances in Neural Information Proc. Systems (NIPS)*, vol. 33, pp. 3846–3857, 2020.

[17] N. Ito and M. Sugiyama, "Audio signal enhancement with learning from positive and unlabeled data," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.

[18] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, "Noisy-target training: A training strategy for dnn-based speech enhancement without clean speech," in *European Signal Proc. Conf. (EUSIPCO)*, Dublin, Ireland, 2021, pp. 436–440.

[19] X. Hao, S. Wen, X. Su, Y. Liu, G. Gao, and X. Li, "Sub-Band Knowledge Distillation Framework for Speech Enhancement," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 2687–2691.

[20] S. Nakaoka, L. Li, S. Inoue, and S. Makino, "Teacher-student learning for low-latency online speech enhancement using Wave-U-Net," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Singapore, Singapore, 2021, pp. 661–665.

[21] H. Fang, D. Becker, S. Wermter, and T. Gerkmann, "Integrating uncertainty into neural network-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 31, pp. 1587–1600, 2023.

[22] H. Fang and T. Gerkmann, "Uncertainty estimation in deep speech enhancement using complex Gaussian mixture models," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.

[23] 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.

[24] K.-L. Chen, D. D. Wong, K. Tan, B. Xu, A. Kumar, and V. K. Ithapu, "Leveraging heteroscedastic uncertainty in learning complex spectral mapping for single-channel speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.

[25] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds., Wiley, 2018, pp. 65–85.

[26] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" *Advances in Neural Information Proc. Systems (NIPS)*, vol. 30, 2017.

[27] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Proc. Systems (NIPS)*, vol. 30, 2017.

[28] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 30, pp. 2993–3007, 2022.

[29] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 676–680.

[30] C. K. Reddy, V. Gopal, R. Cutler, *et al.*, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 2492–2496.

[31] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Sennheiser LDC93S6B," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[32] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, USA, 2015, pp. 504–511.

[33] K. Tan and D. Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 3229–3233.

[34] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 656–660.

2.4 Uncertainty-Driven Hybrid Fusion for Audio-Visual Phoneme Recognition [P5]

Reference

H. Fang, S. Frintrop, and T. Gerkmann, "Uncertainty-driven hybrid fusion for audio-visual phoneme recognition," in *Speech Communication; 15th ITG Conference*, Aachen, Germany, 2023, pp. 255–259. DOI: 10.30420/456164050

Copyright Notice

Uncertainty-Driven Hybrid Fusion for Audio-Visual Phoneme Recognition

Huajian Fang[†1]*, Simone Frintrop*[*]*, Timo Gerkmann*[†]

[†]Signal Processing, [*]Computer Vision, Universität Hamburg, Hamburg, Germany
Email: {huajian.fang, simone.frintrop, timo.gerkmann}@uni-hamburg.de

Abstract

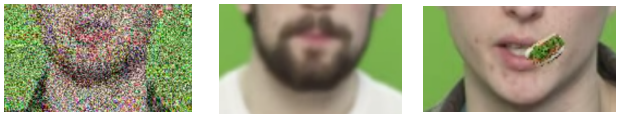
For several speech-processing tasks, complementary features from the visual modality may improve model performance. However, unreliable visual input may provide misleading information, resulting in degraded performance that may be even worse than methods based solely on the audio modality. In this work, we propose an uncertainty-driven hybrid fusion scheme for audio-visual phoneme recognition, mitigating the impact of an unreliable visual modality. More specifically, we incorporate modality-wise uncertainty into decision-making, enabling the model to adaptively determine whether to combine multiple modalities and the extent to which the decision depends on each modality. Experimental results show that the proposed uncertainty-driven hybrid fusion scheme retains the benefits of multi-modal approaches when visual inputs are clean and informative, while at the same time being robust to visual modality distortions.

1 Introduction

Clean speech recorded by a microphone is often corrupted by interfering sounds, which causes difficulties for machines to understand via recognition systems [1]. Phoneme recognition aims to recognize underlying phonetic patterns from corrupted speech signals [2]. However, achieving model robustness across different acoustic distortions has been challenging. Recent work has shown that this problem can be alleviated by incorporating other modalities such as vision, since acoustic noise distortions do not affect the associated visual data [2, 3]. In analogy to how humans use lipreading to aid comprehension in heavily distorted acoustic environments, audio-visual approaches can take advantage of articulation features provided by the visual input (e.g., by processing speakers' mouths) to improve recognition performance at low signal-to-noise ratios (SNRs).

While conventional approaches rely on, e.g., hidden Markov models [2, 4], recent research trends have mainly adopted deep neural networks (DNNs) due to their flexibility and powerful non-linear modeling capacities [5–8]. DNN-based audio-visual recognition pipelines [5–8] often leverage learned features instead of hand-crafted features, which allows for end-to-end training and potentially obtains a task-specific signal transform. Multi-modal approaches involve a necessary step, i.e., modality fusion, playing a crucial role in the resulting performance [3]. DNN-based audio-visual fusion paradigms can be roughly categorized into *early fusion*, which combines different modalities at the input space; *intermediate fusion*, which occurs at a high dimensional latent representation space; *late fusion*, which combines the modalities at the decision level after being separately processed [2, 3]. Although early fusion enables tight integration of different modalities, it is non-trivial to design a single model capable of processing inherently different modalities. Late fusion is a practically easier alternative and allows for independent uni-modal model design. In contrast, intermediate fusion flexibly combines discriminative features extracted from different modalities at an intermediate level, followed by a post-processing module to further integrate their correlations, and has been mostly utilized in the existing systems [3].

While multi-modal methods have demonstrated benefits over uni-modal methods [3, 6, 8, 9], most audio-visual approaches are based on the less-than-realistic assumption that video inputs are consistently clean and informative. Our analysis indicates that under unreliable visual input (as in the examples provided in Figure 1), the recognition performance of the audio-visual approach is even worse than that of the uni-modal approach. A similar finding has also been shown in recent work [10]. Therefore, when the visual modality is too noisy to provide useful information, the recognition system should adaptively rely more on the audio modality, so that the multi-modal approach can maintain the same performance as the audio-only approach. This calls for a model that provides not only the target prediction but also the associated confidence (or uncertainty), according to which a decision can be made on whether to incorporate the corresponding modality.



Additive Noise Blurring Occlusion

Figure 1: Examples of visual distortions with additive Gaussian noise, blurring, and object occlusion.

Predictive uncertainty is typically categorized into *data uncertainty*, which indicates inherent uncertainty in data (e.g., phoneme confusion between /m/ and /n/ for audio inputs, and /b/ and /p/ for visual inputs), and *model uncertainty*, which describes uncertainty in the parameters of a DNN [11]. When we consider phoneme recognition as a classification task, data uncertainty can be computed based on estimated probability scores [12]. In contrast, model uncertainty estimation is often performed by modeling the parameters of a DNN stochastically rather than deterministically, i.e., approximating the posterior distribution of the stochastic network parameters [12]. Among existing approximate Bayesian inference methods, only a few have shown scalability to large DNN models, such as ensemble-based [13] and variational inference-based [14, 15] approaches. For example, Gal et al. [14] perform variational inference and interpret the dropout regularization as imposing Bernoulli distribution on the DNN weights, referred to as *Monte Carlo (MC) dropout*.

In this work, we depart from the assumption that the video input is consistently clean and reliable; instead, we investigate how to improve the model's robustness to simultaneously corrupted video and audio. More specifically, we propose to tackle this problem by incorporating modality-wise uncertainty to adaptively rely more on the audio modality in decision-making when the visual input is distorted and unreliable. At the same time, similar to intermediate fusion, we want to retain the better recognition ability of the multi-modal method when the visual input is clean. For this, we propose an uncertainty-driven hybrid fusion strategy. To achieve that, we first extend the model based on the intermediate fusion [8] to a three-branch model, i.e., the model can perform audio-only, video-only, and audio-visual recognition simultaneously. We further model the predictive uncertainty of each recognition branch, based on which we design a simple yet effective uncertainty-driven hybrid fusion strategy, which leverages both intermediate fusion and late fusion (thus *hybrid fusion*). Our experimental results show that the proposed hybrid fusion strategy is robust to visual corruption and meanwhile capable of incorporating the complementary feature from clean and reliable video to improve phoneme accuracy. Note that, in contrast to previous works that include visual input corruption during training [10, 16], our uncertainty-driven hybrid fusion is *video corruption-agnostic*, i.e., our video model is only trained on clean video data, which is expected to generalize widely to different types of video distortions.

[1]This work has been funded by ahoi.digital.

2 Audio-Visual Phoneme Recognition

In this work, we focus on DNN-based audio-visual phoneme recognition, where we formulate the problem as a typical classification problem as in [8]. Specifically, the proposed hybrid fusion scheme is built on top of the widely-used intermediate feature fusion scheme [6, 8], which consists of two separate models processing audio and video, followed by a multi-modal feature fusion model.

An audio-only phoneme recognition model consists of learnable transform and feature processing modules, which extract relevant acoustic features and refine these features into discriminative phonetic representations. Equivalent to phonemes inferred from speech sounds, visemes are defined based on the appearance of lips when articulating phonemes [17]. However, multiple phonemes can be grouped into a viseme category, that is, the correspondence between phonemes and visemes is a many-to-one mapping. This makes visual speech recognition inferring phonemes from visual data only an inherently difficult task. Nevertheless, recent advances have shown that carefully-designed deep learning-based methods, especially in combination with temporal modeling, can yield promising results [5, 7, 8].

As data from different modalities may be recorded at different sampling rates, synchronization is another consideration when dealing with multiple modalities, i.e., audio and video inputs needs to be temporally aligned. As in this work phonemes are recognized at a frame level, the frame rate of audio signals depends on the frame size selected; the video frame rate is equipment-related. Temporal alignment is also application-oriented. For example, a recent audio-visual method for speech enhancement has applied pooling techniques to the audio input to match the temporal resolution of the visual stream in the high-dimensional latent space [9]. In this work, we address this problem by upsampling the video input to match the frame rate of the audio signal.

With the synchronized representative audio and visual features in the latent space, the fusion block aims to exploit the correlations between the audio and visual modalities. To effectively combine different modalities, various fusion strategies have been proposed, such as attention-based fusion [18], squeeze-excitation fusion [19], and addition- and concatenation-based fusion [3]. However, generalizing a fusion technique across different datasets and tasks remains challenging and its performance is often architecture- and task-dependent. Nevertheless, the concatenation-based fusion strategy is often preferred in multi-modal methods, partially due to its simplicity of implementation. Furthermore, a recent empirical comparison of various fusion strategies by Richter et al. [8] has also revealed its effectiveness in phoneme recognition. Thus, we concatenate the audio and visual features along the feature channel as input to the fusion model.

While current DNN-based methods have shown success in the speech recognition task, the generalization ability to unseen inputs is not guaranteed, especially when the model is processing out-of-distribution samples under-represented by training data. Therefore, it is essential to estimate predictive uncertainty in addition to the target prediction. This is particularly useful in audio-visual phoneme recognition because modality-wise uncertainty can help determine how much confidence we can put into the model's prediction without having access to ground truth. Moreover, most existing audio-visual methods are based on the assumption that the visual input is consistently reliable and informative, which is less than realistic, as multiple factors can lead to unreliable visual input, such as object occlusion, data transmission failure, device issues, and illumination conditions. As we will show in the experimental evaluation, misleading visual inputs can cause significant performance decline in the audio-visual models considered. Therefore, when visual input data are unreliable and provide misleading information, the system should adaptively rely on the audio modality only. To achieve this, we propose to make use of uncertainty modeling, which enables the model to output not only target predictions but also associated uncertainty estimates. With this, we aim to improve the model's robustness to visual corruption, while retaining the benefits of multi-modal approaches when the visual input is reliable.

3 Predictive Uncertainty Estimation

Uncertainty estimation in DNN-based methods is a challenging task, where network models typically involve millions of parameters or more [11, 20]. In this work, we estimate the modality-wise uncertainty for audio and video using MC dropout, due to its effectiveness and scalability to large DNN models, as shown in different tasks in previous work, including computer vision tasks semantic segmentation and depth regression [21], as well as speech enhancement [22]. MC dropout establishes a connection between a widely-used regularization technique, dropout [23], and approximating the posterior distribution of the weights of a neural network. Gal et al. provided a detailed derivation in [14]. By activating dropout at testing, we perform multiple stochastic forward passes for each input, simulating the sampling process from the posterior of the weights of a DNN. Consequently, we can obtain a set of softmax probability scores $\{\boldsymbol{p}_m\}_{m=1}^{M}$ for each input frame, where m indexes M sampling times. We can compute the entropy of the expected distribution [12, 14, 24] to approximate total uncertainty, as it takes into account both data uncertainty and model uncertainty. Hereafter, it will be referred to as predictive uncertainty:

$$\mu_c = \frac{1}{M}\sum_{m}^{M} \boldsymbol{p}_{m,c} \quad \text{and} \quad U = -\sum_{c} \mu_c \log(\mu_c), \tag{1}$$

where $\boldsymbol{p}_{m,c}$ indicates the probability score of the c-th class at the m-th forward pass. Note that the predictive uncertainty can be estimated for both uni-modal models (i.e., audio-only and video-only models), and multi-modal models (i.e., audio-visual model). Furthermore, this selected measure of uncertainty is bounded in the sense that a uniform distribution across C classes leads to the largest uncertainty. Thus, it can be further normalized into the range of $[0, 1]$ for intuitive interpretation. Since there is no guarantee that the visual input always provides useful information, here we propose to tackle this problem by leveraging properly captured modality-wise uncertainty, as will be explained next.

4 Uncertainty-Driven Hybrid Fusion

To integrate uncertainty into the framework, we aim to achieve an uncertainty-based fusion scheme, which incorporates the visual modality when it can provide complementary features to the audio input, and relies more on the audio modality when the visual input is distorted [10, 25]. This requires estimating uni-modal uncertainty in the intermediate fusion scheme to quantify the confidence in each modality. However, most existing algorithms performing the feature fusion in the latent space do not allow for uni-modal phoneme recognition, as only the fusion model is tasked to output probabilities based on the correlations of audio-visual inputs. Therefore, we first extend the general intermediate fusion model to a three-branch model, where besides concatenating the audio-visual features and feeding them into the fusion model, we also keep the uni-modal classifier that performs uni-modal phoneme classification. The proposed framework is illustrated in Figure 2. The new design does not introduce a large computational overhead compared with the general intermediate fusion model, as only a fully-connected layer is added for each modality to output uni-modal probability scores. Eventually, the proposed three-branch fusion scheme can perform end-to-end training by a joint loss:

$$L = \beta L_{\mathrm{a}} + \beta L_{\mathrm{v}} + \alpha L_{\mathrm{f}}; \quad \text{s.t. } \alpha + \beta + \beta = 1, \tag{2}$$

where L_{a}, L_{v}, and L_{f} are the cross entropy losses for the audio, video, and fusion branches, respectively. α and β are hyperparameters that balance the contribution of each branch. After that, we can compute predictive uncertainty measures for each branch prediction as in (1), denoted as U_{a}, U_{v}, and U_{f} respectively. The video uncertainty U_{v} here is computed by first grouping the estimated phoneme probability scores into the corresponding viseme class according to the mapping provided in [8, Table I][26].

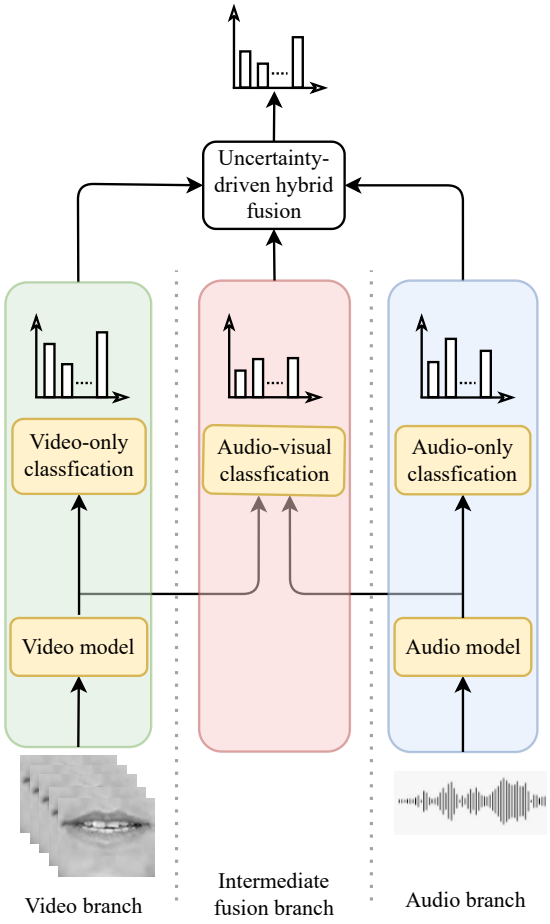


Figure 2: Proposed end-to-end three-branch audio-visual phoneme recognition system. The method takes as input video frames and raw speech waveforms. The uncertainty-based hybrid fusion strategy is performed based on the probability scores of the audio-visual branch (intermediate fusion) and the two uni-modal branches (late fusion).

When the video modality is corrupted, the visual branch prediction is expected to be uncertain (and likely inaccurate), which will be reflected by high uncertainty measures. Consequently, the uncertainty of the fusion branch may also be raised due to possible misleading representations of the video. However, visual corruption may not affect the audio branch, which should play a more important role in decision-making. Therefore, for the case of uncertain video input, we design a weighting strategy between the *audio* and *fusion* branches. We discard the video branch in this case because the video recognition branch is susceptible to visual distortions and has a much worse performance than the intermediate fusion branch, as will be shown later. Furthermore, we do not rely exclusively on the audio branch either, as we still want to rely more on the fusion branch to leverage the benefits of multi-modal methods when the video modality is intact and can provide complementary features. In the case of clean video, the fusion branch that incorporates multiple modalities may exhibit greater certainty than either uni-modal approach. Thus, the weighting strategy needs to adaptively assign a large weight to the fusion branch in this case.

Eventually, we need to determine an uncertainty threshold for the video modality, T_{v}, to indicate whether we incorporate visual information. Since it is still an open question in the uncertainty modeling literature how to define a threshold to distinguish certain from uncertain estimates, here we follow [27] and select the threshold based on the average uncertain value of the *validation* dataset. Besides, we require a strategy so that the audio and fusion branches are weighted according to their uncertainty estimates,

Algorithm 1 Uncertainty-driven hybrid fusion

Input: video uncertainty U_{v}, video uncertainty threshold T_{v}, audio uncertainty U_{a}, audio score P_{a}, fusion uncertainty U_{f}, intermediate fusion score P_{f}
if $U_{\mathrm{v}} < T_{\mathrm{v}}$ **then**
  $P_{\mathrm{h}} = P_{\mathrm{f}}$
else
  $W_{\mathrm{a}},\ W_{\mathrm{f}} = f(U_{\mathrm{a}},\ U_{\mathrm{f}})$ ▷ $W_{\mathrm{a}} + W_{\mathrm{f}} = 1$
  $P_{\mathrm{h}} = W_{\mathrm{a}} \times P_{\mathrm{a}} + W_{\mathrm{f}} \times P_{\mathrm{f}}$
end if
return hybrid fusion score P_{h}

formulated as $W_{\mathrm{a}},\ W_{\mathrm{f}} = f(U_{\mathrm{a}},\ U_{\mathrm{f}})$, where W_{a} and W_{f} refer to the weights of the probability scores of the audio and fusion branches returned by the score calculation function $f(*,*)$. While there are various possible solutions to define the weighting score function, here we provide a simple strategy, defined as:

$$R_{\mathrm{a}} = 1 - U_{\mathrm{a}}, \quad R_{\mathrm{f}} = 1 - U_{\mathrm{f}}$$
$$W_{\mathrm{a}} = \frac{R_{\mathrm{a}}}{R_{\mathrm{a}} + R_{\mathrm{f}}}, \quad W_{\mathrm{f}} = \frac{R_{\mathrm{f}}}{R_{\mathrm{a}} + R_{\mathrm{f}}}. \tag{3}$$

The uncertainty-driven hybrid fusion strategy is summarized in Algorithm 1. By incorporating uncertainty into decision-making, the method aims to adaptively fuse the multi-modal outputs, improving the model's robustness to visual distortions.

5 Experimental Setup

5.1 Data

In this work, we use the publicly available dataset NTCD-TIMIT [28], which is a noisy version of TCD-TIMIT [29], which is, in turn, an audio-visual version of TIMIT [30]. The speech material in the corpus is split into approximately 5 hours (17 speakers), 1 hour (8 speakers), 1 hour (9 speakers) for training, validation, and testing, respectively. The NTCD-TIMIT has been created by mixing speech utterances with 6 types of acoustic noise: white, babble, car, living room, street, and cafe, at 6 different SNRs: {-5, 0, 5, 10, 15, 20} dB. To ensure that the model is tested on completely unseen acoustic noise, the clean speech test set is mixed with the noise signals from the QUT corpus [31], at the same range of SNRs. All audio signals are sampled at 16 kHz. We consider 38 phoneme classes as in [8, Table I][4].

To test the robustness of the audio-visual model to unreliable visual input, we simulate the same video corruption as in [10] [1], including occluding a speaker's mouth with Naturalistic Occlusion Generation (NatOcc) patches from [32], blurring video frames, and introducing additive Gaussian noise. To corrupt the mouth region of interest, it is necessary to perform face and landmark detection, where we follow the preprocessing pipeline in [33]. The cropped visual input frames of size 67×67 are converted to grayscale for computational efficiency. Note that video corruption only occurs during testing and has not been included in the training.

5.2 Architecture and hyperparameters

For the architecture, we use the audio-visual backbones in [8] [2]. For the audio model, we use a 1-D convolutional layer followed by ResNet-18 [34] and two bi-directional gated recurrent unit (BGRU) layers with dropout layers (dropout rate 0.5) inserted before and after the first BGRU layer. The video model is based on three 3-D convolutional layers followed by two BGRU layers used in [8, 35]. The fusion model consists of two layers of BGRU (512 units) followed by a fully-connected output layer with the output

[1] https://github.com/joannahong/AV-RelScore
[2] https://github.com/sp-uhh/av-phoneme

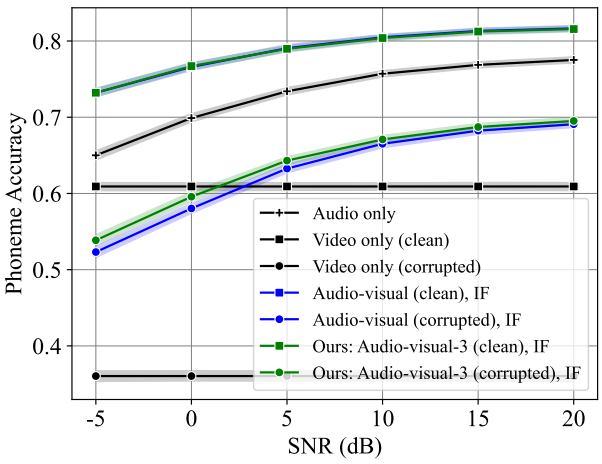


Figure 3: Phoneme accuracy of the synthetic QUT-TIMIT test set. The line plot represents the mean with a 95% confidence interval. Our proposed three-branch multi-modal model is referred to as "*Audio-visual-3*" and "*IF*" indicates the intermediate fusion branch of "*Audio-visual-3*". "*Audio-visual, IF*" indicates the audio-visual baseline model using only the intermediate fusion scheme. *(clean)* and *(corrupted)* indicate the clean and corrupted input video. Note that "*Audio-visual (clean), IF*" and "*Ours: Audio-visual-3 (clean), IF*" are visually overlapping.

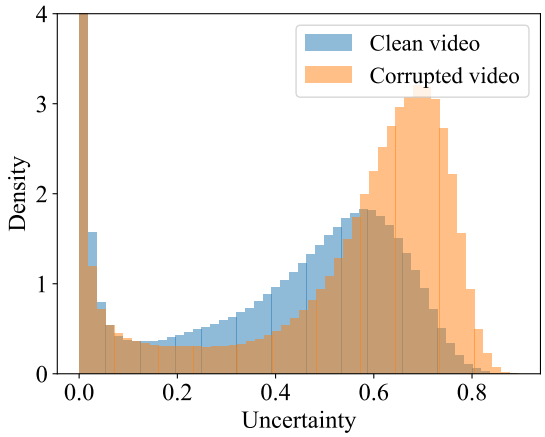


Figure 4: Uncertainty distribution of the video branch in *Audio-visual-3* with clean and corrupted video inputs. The density is computed based on the time frames in the test set.

dimension $C = 38$. Each uni-modal classifier in Figure 2 is a fully-connected layer. Both audio and visual models provide features of dimension 512 for each frame. Since we perform phoneme recognition at a frame level, the audio model is designed to obtain a frame rate of 62.5 Hz (similar to a deterministic transform with a window size of 64 ms and 75% overlap). The video input is upsampled to the same frame rate using the FFmpeg tool [36]. We optimize the DNN using Adam optimizer with an initial learning rate of 10^{-3}. We set the batch size to 16; the training is early stopped with a patience of 10 epochs; α is set to 0.9.

6 Results

In this work, we compare the proposed three-branch model, denoted as *Audio-visual-3*, with the baseline audio-visual model using only the intermediate fusion scheme [8], denoted as *Audio-visual, IF*.

We present the phoneme accuracy of the uni-modal and multi-modal methods as a function of the input audio SNR in Figure 3. Our results first confirm the benefits of the audio-visual approach that leverages complementary features compared to the uni-modal approach. For our proposed three-branch model, the performance

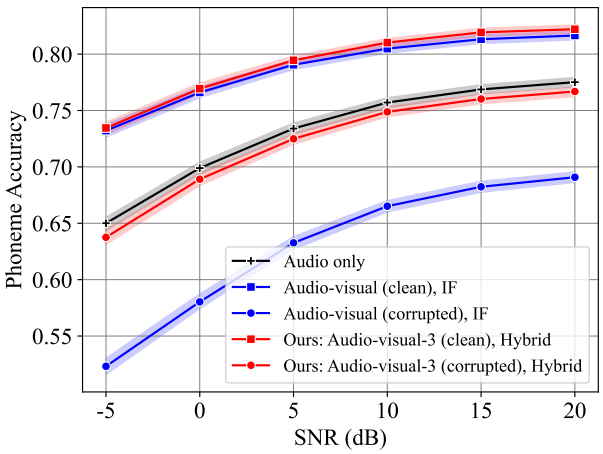


Figure 5: Phoneme accuracy of the synthetic QUT-TIMIT test set (we compare the proposed hybrid scheme only with the relevant approaches in Figure 3). *Hybrid* indicates the proposed hybrid fusion strategy. Please refer to the caption of Figure 3 for other naming conventions.

of the middle fusion branch is reported here for a fair comparison with the baseline intermediate fusion scheme. With the corrupted video input, the performance of the video-only model declines significantly; similar behavior can be observed for both the baseline audio-visual method (*Audio-visual (corrupted), IF*) and the intermediate fusion branch of the proposed three-branch model (*Audio-visual-3 (corrupted), IF*). This drop is expected, as unreliable visual information may mislead the fusion model and cause performance degradation.

Next, we take the video branch of *Audio-visual-3* and analyze its predictive uncertainty distribution with and without video corruption, as shown in Figure 4. It can be observed that the video model provides larger uncertainties for corrupted inputs, indicating that the model provides reliable uncertainty estimates for unseen and insufficiently represented input samples. By taking this estimated predictive uncertainty into account as in Algorithm 1, the performance of our proposed hybrid fusion scheme (*Audio-visual-3 (corrupted), Hybrid*) largely outperforms the baseline (*Audio-visual (corrupted), IF*) and is very close to that of the audio-only model when the corrupted video is present, as shown in Figure 5. This demonstrates that the proposed uncertainty-driven fusion scheme relies more on the audio modality when the visual input becomes less instructive. At the same time, we can observe that it performs comparably to the baseline audio-visual model on clean video input, indicating that the proposed uncertainty-driven scheme is capable of incorporating complementary features of the clean video.

7 Conclusion

In this work, we present an uncertainty-driven hybrid fusion scheme to alleviate the impact of unreliable video inputs. By considering modality-wise uncertainty as a reliability indication of the prediction, we integrate it into decision-making for audio-visual phoneme recognition. The proposed hybrid fusion strategy has demonstrated its improved robustness to unseen visual corruption compared to the baseline audio-visual method that only uses the intermediate fusion scheme, while retaining the benefits offered by multi-modal methods when the visual input is informative. Future work may include exploring more sophisticated architectures and how to extract useful information from partially distorted videos to further improve performance on corrupted visual inputs.

References

[1] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement*,

pp. 65–85, Wiley, 2018.

[2] A. H. Abdelaziz, "Comparing fusion models for DNN-based audiovisual continuous speech recognition," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 26, no. 3, pp. 475–484, 2018.

[3] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 29, pp. 1368–1396, 2021.

[4] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, vol. 37, no. 11, pp. 1641–1648, 1989.

[5] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in *INTERSPEECH*, pp. 3652–3656, 2017.

[6] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, pp. 7613–7617, 2021.

[7] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, pp. 1–10, 2022.

[8] J. Richter, J. Liebold, and T. Gerkamnn, "Continuous phoneme recognition based on audio-visual modality fusion," in *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2022.

[9] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 15490–15500, IEEE, 2021.

[10] J. Hong, M. Kim, J. Choi, and Y. M. Ro, "Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 18783–18794, 2023.

[11] 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.

[12] A. Malinin, *Uncertainty estimation in deep learning with application to spoken language assessment*. PhD thesis, University of Cambridge, 2019.

[13] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Proc. Systems (NIPS)*, vol. 30, 2017.

[14] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Int. Conf. Machine Learning (ICML)*, pp. 1050–1059, June 2016.

[15] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Int. Conf. Machine Learning (ICML)*, pp. 1613–1622, June 2015.

[16] T. Afouras, J. S. Chung, and A. Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," in *Interspeech*, pp. 4295–4299, 2019.

[17] H. L. Bear and R. Harvey, "Phoneme-to-viseme mappings: the good, the bad, and the ugly," *Speech Communication*, vol. 95, pp. 40–67, 2017.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Proc. Systems (NIPS)*, vol. 30, 2017.

[19] M. L. Iuzzolino and K. Koishida, "Av (se) 2: Audio-visual squeeze-excite speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, pp. 7539–7543, 2020.

[20] H. Fang and T. Gerkmann, "Uncertainty estimation in deep speech enhancement using complex gaussian mixture models," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, pp. 1–5, 2023.

[21] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," *Advances in Neural Information Proc. Systems (NIPS)*, vol. 30, 2017.

[22] H. Fang, D. Becker, S. Wermter, and T. Gerkmann, "Integrating uncertainty into neural network-based speech enhancement," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 31, pp. 1587–1600, 2023.

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[24] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 605–613, Springer, 2019.

[25] M. Subedar, R. Krishnan, P. L. Meyer, O. Tickoo, and J. Huang, "Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference," in *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pp. 6301–6310, October 2019.

[26] A. Markides, "Speechreading (lipreading).," *Child: care, health and development*, 1979.

[27] J. Mukhoti and Y. Gal, "Evaluating bayesian deep learning methods for semantic segmentation," *arXiv preprint arXiv:1811.12709*, 2018.

[28] A. H. Abdelaziz *et al.*, "NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition.," in *Interspeech*, pp. 3752–3756, 2017.

[29] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[30] J. S. Garofolo, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[31] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Interspeech*, pp. 3110–3113, 2010.

[32] K. T. Voo, L. Jiang, and C. C. Loy, "Delving into high-quality synthetic face occlusion segmentation datasets," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4711–4720, 2022.

[33] P. Ma, Y. Wang, S. Petridis, J. Shen, and M. Pantic, "Training strategies for improved lip-reading," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, pp. 8472–8476, 2022.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2016.

[35] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "Lipnet: End-to-end sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, 2016.

[36] S. Tomar, "Converting video formats with FFmpeg," *Linux journal*, vol. 2006, no. 146, p. 10, 2006.

CHAPTER 3

Noise-Aware Generative Speech Enhancement Based on Variational Autoencoder and Non-Negative Matrix Factorization

3.1 Variational Autoencoder for Speech Enhancement with a Noise-Aware Encoder [P6]

Reference

H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, ON, Canada, 2021, pp. 676–680. DOI: 10.1109/ICASSP39728.2021.9414060

Copyright Notice

VARIATIONAL AUTOENCODER FOR SPEECH ENHANCEMENT WITH A NOISE-AWARE ENCODER

Huajian Fang[1,2]*, Guillaume Carbajal*[1]*, Stefan Wermter*[2]*, Timo Gerkmann*[1]

[1]Signal Processing (SP), Universität Hamburg, Germany
[2]Knowledge Technology (WTM), Universität Hamburg, Germany
{fang, carbajal, wermter, gerkmann}@informatik.uni-hamburg.de

ABSTRACT

Recently, a generative variational autoencoder (VAE) has been proposed for speech enhancement to model speech statistics. However, this approach only uses clean speech in the training phase, making the estimation particularly sensitive to noise presence, especially in low signal-to-noise ratios (SNRs). To increase the robustness of the VAE, we propose to include noise information in the training phase by using a *noise-aware encoder* trained on noisy-clean speech pairs. We evaluate our approach on real recordings of different noisy environments and acoustic conditions using two different noise datasets. We show that our proposed noise-aware VAE outperforms the standard VAE in terms of overall distortion without increasing the number of model parameters. At the same time, we demonstrate that our model is capable of generalizing to unseen noise conditions better than a supervised feedforward deep neural network (DNN). Furthermore, we demonstrate the robustness of the model performance to a reduction of the noisy-clean speech training data size.

***Index Terms*—** speech enhancement, generative model, variational autoencoder, semi-supervised learning.

1. INTRODUCTION

Speech enhancement refers to the problem of extracting a target speech signal from a noisy mixture in order to enhance the quality and intelligibility of the speech. This task is of particular interest for applications like speech recognition and hearing aids. Single-channel speech enhancement is a challenging task, especially at low signal-to-noise ratios (SNRs).

Speech enhancement typically requires the statistical estimation of the noise and speech power spectral densities (PSDs) [1, 2]. Non-negative matrix factorization (NMF) is a popular choice for PSD estimation [3–6]. However, underlying linearity assumptions limit the performance when modeling complex high-dimensional data. In contrast, speech enhancement based on non-linear deep neural networks (DNNs) has shown better modeling capacity. Common approaches focus on inferring a time-frequency mask in a supervised manner [7]. However, to generalize to unseen noise conditions, DNNs require a large number of pairs of noisy and clean speech in various acoustic conditions [8].

Recently, there has been an increasing interest in generative models, such as generative adversarial networks (GANs) [9] and variational autoencoders (VAEs) [10, 11]. The generative VAE is a probabilistic model widely used for learning latent representations of a probabilistic distribution. The VAE features a similar architecture as a classical autoencoder with an encoder and a decoder, but its latent space differs by being regularized to follow a standard Gaussian distribution. Moreover, the VAE has been extended to deep conditional generative models for effectively performing probabilistic inference [12,13]. VAEs have been applied to speech enhancement in both single-channel and multi-channel scenarios [14–16]. They have been used to model the speech statistics by training on clean speech spectra only. However, because no noise information is involved in its training phase, the encoder of the standard VAE is sensitive to noise. In low SNRs, this noise-sensitivity results in the erroneous estimation of latent variables and thus in inappropriately generated speech coefficients and a reduced performance.

In this work, inspired by conditional VAEs and its application to image segmentation [12, 13, 17], to increase noise robustness, we propose to replace the encoder of the VAE by a *noise-aware encoder*. To learn this encoder, the VAE is first trained on clean speech spectra only, and then, given noisy speech, the proposed noise-aware encoder is trained in a supervised fashion to make its latent space as close as possible to that of the first speech-only trained encoder. For our analyses we rely on the VAE-NMF speech enhancement framework [14, 15], which uses NMF to model the noise PSD. We show that the proposed encoder is more robust to noise presence and improves speech estimation without increasing the number of model parameters. The method also shows robustness to unseen noise conditions by evaluating on real recordings from different noise datasets. Finally, we illustrate that already a small amount of noisy-clean speech data can lead to improvements in overall distortion.

In section 2, we introduce problem settings and notations, as well as the framework of the VAE-based speech model and the noise model developed on the NMF. In section 3, we introduce details about the proposed noise-aware VAE. After showing the experiment settings in section 4, we present experimental evaluation results and conclusions in section 5 and section 6.

2. PROBLEM FORMULATION

2.1. Mixture model

In our work, we employ an additive signal model, where a noisy mixture is seen as a superposition of clean speech and additive noise. In the short-time Fourier transform (STFT) domain, it shows as

$$x_{ft} = s_{ft} + n_{ft}, \tag{1}$$

where x_{ft}, s_{ft}, and n_{ft} represent each time-frequency coefficient in spectra of noisy mixture $X \in \mathbb{C}^{F\times T}$, speech $S \in \mathbb{C}^{F\times T}$, and noise $N \in \mathbb{C}^{F\times T}$ respectively. F denotes the number of frequency bins, T represents the number of time frames, which are indexed by f and t, respectively. The speech and noise spectra are assumed to be mutually independent complex Gaussian distributions with zero-mean, i.e., $s_{ft} \sim \mathcal{N}_{\mathbb{C}}(0,\, \sigma^2_{s,ft}), n_{ft} \sim \mathcal{N}_{\mathbb{C}}(0,\, \sigma^2_{n,ft})$ where $\sigma^2_{s,ft}$,

$\sigma^2_{n,ft}$ represent the variances of speech and noise. The PSD of signals is characterized by the parameter variance under the local stationary assumption [18].

Furthermore, to provide an increased robustness to the loudness of the audio utterances, a time-dependent and frequency-independent gain g_t is introduced [15]. Eventually, this modifies the additive mixture model in (1) to

$$x_{ft} = \sqrt{g_t} s_{ft} + n_{ft}. \tag{2}$$

Given the observed noisy mixture which follows a complex Gaussian distribution as $x_{ft} \sim \mathcal{N}_{\mathbb{C}}(0,\, g_t \sigma^2_{s,ft} + \sigma^2_{n,ft})$, the desired speech can be extracted by separately modeling the speech and noise variances.

2.2. Speech model

For the VAE-based speech model, a frame-wise D-dimensional latent variable $z_t \in \mathbb{R}^D$ is defined, and an F-dimensional speech frame s_t is assumed to be sampled from the conditional likelihood distribution $p_\theta(s_t|z_t)$. This is achieved by the decoder of VAE, also called the generative model. The variable θ here indicates the parameters of the decoder network. $\hat{\sigma}^2_s : \mathbb{R}^D \rightarrow \mathbb{R}^F_+$ denotes the nonlinear function from the latent space to the reconstructed signal given by the generative model of the VAE.

The VAE provides a principled method to jointly learn latent variables and the inference model [10]. Following a Bayesian framework, this requires to approximate the intractable true posterior distribution $p(z_t|s_t)$. In the VAE, the encoder, also called the inference model, is used to approximate the true posterior, denoted as $q_\phi(z_t|s_t)$. The variable ϕ here indicates the parameters of the encoder network. $\hat{\mu}_d : \mathbb{R}^F_+ \rightarrow \mathbb{R}^D$, $\hat{\sigma}^2_d : \mathbb{R}^F_+ \rightarrow \mathbb{R}^D_+$ indicate the nonlinear mapping of the neural network given by the inference model of the VAE. Under stochastic gradient descent, the generative model's parameters θ and the inference model's parameters ϕ are jointly optimized by maximizing variational lower bound, given by

$$\begin{aligned} \log p(S) \geq & -\sum_t \mathbb{KL}[q_\phi(z_t|s_t)||p(z_t))] \\ & + \sum_t \mathbb{E}_{q_\phi(z_t|s_t)}[\log p_\theta(s_t|z_t)]. \end{aligned} \tag{3}$$

The quantity $p(z_t)$ represents the prior distribution of the D-dimensional variable z_t, and $\mathbb{KL}$ indicates Kullback-Leibler divergence. The prior of the latent variables is defined as a zero-mean isotropic multivariate Gaussian $z_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as in [10]. The first term in the objective function (3) refers to the regularization error in the latent space to ensure meaningful latent variables, and the second term is the reconstruction error.

As shown in Fig. 1, the VAE is trained on the periodograms of clean speech $|s_t|^2$ [14, 15]. During testing, the estimates of the clean speech power spectra $\hat{\sigma}^2_s(z_t)$ are expected to be generated from latent variables learnt from the noisy periodograms $|x_t|^2 \in \mathbb{R}^F_+$. Note that a robust estimation of latent variables that represents the clean speech statistics plays a crucial role in the generative process.

2.3. Noise model

NMF tries to find an optimal approximation to an input matrix by a dictionary matrix containing basis functions weighted by a coefficients matrix [3]. Here NMF is used to model the noise variance [14, 15]. The variance of noise σ^2_n is approximated by a multiplication of the dictionary matrix $W \in \mathbb{R}^{F\times K}_+$ and the coefficients matrix $H \in \mathbb{R}^{K\times T}_+$, computed as

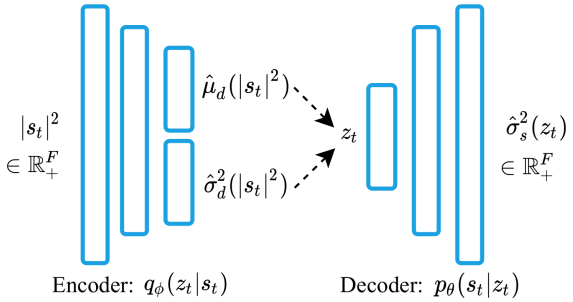


Fig. 1. The generative model and inference model of the adopted VAE. The dashed line here indicates the sampling process.

$$\sigma^2_n = WH = \sum_{ft}\sum_k w_{fk} h_{kt}, \tag{4}$$

where K indicates the rank of the noise model indexed by k. w_{fk} and h_{kt} are elements from W and H respectively at the corresponding row and column indexed by f, k, and t.

2.4. Clean speech inference

By modeling speech and noise with VAE and NMF respectively, the distribution of the noisy mixture can be represented as

$$x_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, g_t \hat{\sigma}^2_{s,f}(z_t) + \sum_k w_{fk} h_{kt}), \tag{5}$$

where $\hat{\sigma}^2_{s,f} : \mathbb{R}^D \rightarrow \mathbb{R}_+$ denotes the nonlinear function $\hat{\sigma}^2_s$ for f-th frequency bin. Given the noisy mixture as an observation, the Monte Carlo expectation-maximization (MCEM) algorithm is utilized to estimate the NMF parameters and the gain factor [15, 19]. The sampling strategy is based on the Metropolis-Hastings algorithm [20]. The clean speech can be extracted from a noisy mixture in the time-frequency domain by constructing a Wiener filter denoted by $\hat{m}_{ft}$, given as

$$\hat{m}_{ft} = \frac{\hat{\sigma}^2_{s,f}(z_t)}{g_t \hat{\sigma}^2_{s,f}(z_t) + \sum_k w_{fk} h_{kt}}. \tag{6}$$

Although modeling speech with a VAE can be achieved by training solely on clean speech data, using it for speech enhancement is another matter since gaining robustness to noise is difficult without including noise samples in the training data and the model. However, the standard VAE does not allow for including noise at the training phase.

3. NOISE-AWARE VAE

Instead of using the encoder trained on the clean speech signals, we propose a noise-aware VAE that can improve the robustness of the encoder against noise presence. For a generative process, it is difficult or even impossible to derive the optimal mapping between latent variables and targets. However, we argue that it might be relevant to make latent variables estimated from noisy mixtures as close as possible to the ones inferred from the corresponding clean speech.

To obtain the noise-aware VAE based on this assumption, we propose a two-step learning algorithm, which learns a non-linear mapping from the noisy signals to latent variables that represent the clean speech statistics. We first train a VAE using Equation (3) to learn a regularized

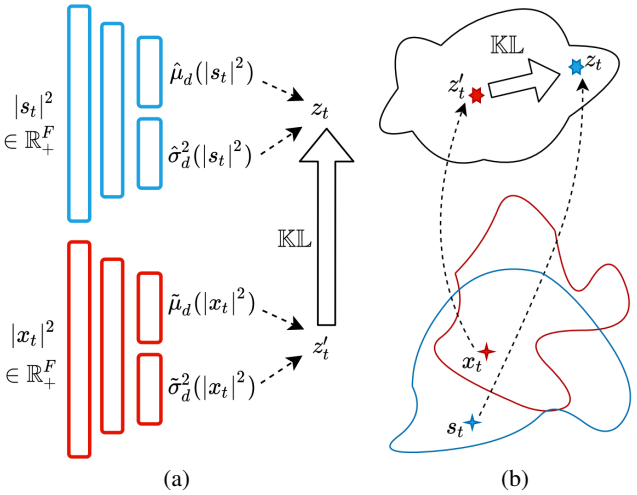


Fig. 2. The proposed architecture for minimizing divergence between latent variables. The constraint in the latent space is shown in (a), and its graphic explanation given in (b).

latent space over the clean speech signals. The noise-aware encoder is then proposed to approximate the probability $q_\gamma(z'_t|x_t)$ to output D-dimensional latent variables $z'_t \in \mathbb{R}^D$ conditioned on the noisy mixture x_t. It is also assumed that the conditional probability $q_\gamma(z'_t|x_t)$ follows a standard Gaussian distribution. The variable γ indicates the parameters of the new encoder. Finally, the distance of z'_t obtained from noisy speech to the latent variables z_t inferred form the corresponding clean speech is minimized based on the Kullback–Leibler divergence as shown in Fig. 2 (a), given by

$$\mathcal{L}(\gamma) = \sum_t \mathbb{KL}(q_\phi(z_t|s_t)||q'_\gamma(z'_t|x_t)) \tag{7}$$

$$= \sum_{t,d} \Big\{ \frac{1}{2} \log \frac{\tilde{\sigma}_d^2(|x_t|^2)}{\hat{\sigma}_d^2(|s_t|^2)} - \frac{1}{2} + \frac{\hat{\sigma}_d^2(|s_t|^2) + (\hat{\mu}_d(|s_t|^2) - \tilde{\mu}_d(|x_t|^2))^2}{2\tilde{\sigma}_d^2(|x_t|^2)} \Big\} \tag{8}$$

where $\tilde{\mu}_d : \mathbb{R}_+^F \to \mathbb{R}^D$ and $\tilde{\sigma}_d^2 : \mathbb{R}_+^F \to \mathbb{R}_+^D$ represents the nonlinear mapping of the neural networks for the mean and variance of the posterior Gaussian distribution for the variable z'_t. The parameters of the new inference model γ are optimized by minimizing the cost function using stochastic gradient descent algorithms. In this way, we combine unsupervised learning of the speech characteristics by the VAE and supervised learning using the pairs of noisy-clean speech signals.

Eventually, as graphically shown in Fig. 2 (b), by introducing this cost function in the latent space, the latent variables z'_t estimated from the noisy mixture x_t is pulled towards z_t estimated from the corresponding clean speech s_t. The dashed lines here indicate the nonlinear mapping from the signal space to the latent space, and different colors indicate two mapping pairs. At the inference stage, the noise-aware inference model is used to replace the standard speech-based encoder. The decoder of the VAE remains unchanged.

4. EXPERIMENTAL SETTINGS

4.1. Datasets

We evaluate the performance of the proposed model by using signals from the speech dataset Wall Street Journal (WSJ0) [21], and the noise databases QUT-NOISE [22] and DEMAND [23]. QUT-NOISE is used in constructing datasets of both training and evaluation using 4 noise types "cafe", "car", "home", and "street" recorded in unique locations. DEMAND is introduced as another evaluation dataset corresponding to completely unseen noise conditions in the training set, and the noise signals are randomly sampled from recordings of 12 noise types in the categories "domestic", "public", "street", and "transportation".

To train the noise-aware encoder, around 25 hours of speech samples are chosen from WSJ0 and mixed with the sampled noise signals at a SNR randomly chosen from the range of -5 dB to 5 dB with a gap of 1 dB. Two speaker-independent evaluation datasets each containing around 2.3 hours of 1000 noisy samples are created by mixing the speech and noise signals at SNRs of -10 dB, -5 dB, 0 dB, 5 dB, and 10 dB.

4.2. Baselines

We show evaluation results by comparing the proposed noise-aware VAE to the standard VAE, and a fully-connected DNN model. The DNN model outputs a Wiener filter based on a mean square error cost function [24], referred to as DNN-WF. The standard VAE is trained on the same amount of the clean speech signals that are not mixed with the noise signals, while the supervised DNN-WF is trained on the same dataset as the noise-aware encoder.

4.3. Hyperparameters

All signals are sampled at 16 kHz. The signal is transformed into the STFT domain with a sine window of length 1024 ($F = 513$) and a 25% hop size. Global normalization to zero mean and unit standard deviation is employed for training the noise-aware encoder, since Kullback–Leibler divergence is scale-dependent. The rank of NMF is chosen to be $K = 8$ when modeling noise, and its composing matrices W and H are randomly initialized. The parameters of MCEM algorithm follow the setting in [15].

The VAE is comprised of an encoder and a decoder both with two feedforward hidden layers of 128 units. The hyperbolic tangent activation function is applied to all hidden layers, except the output layer. The dimension of the latent space L is fixed at 16. The noise-aware encoder has the same structure as the speech-based encoder of the standard VAE. The fully supervised DNN-WF contains 5 hidden layers, each with 128 units, and its architecture is built to contain a similar number of parameters as our VAE model. No temporal information is considered in DNN-WF, which is consistent with the non-sequential characteristic of the VAE. We apply the ReLU activation function to all hidden layers, and the sigmoid function is put on the output layer to ensure the estimate of the Wiener filter mask lies in the range $[0, 1]$. The parameters θ and ϕ of the VAE are optimized by Adam [25] with a learning rate of 1e-3, and the parameters γ of the noise-aware encoder with a learning rate of 1e-4.

4.4. Evaluation metrics

To show the enhancement performance, we employ scale-invariant signal-to-distortion ratio (SI-SDR) in decibel (dB) [26] to measure the overall distortion, which takes both noise reduction and artifacts into account.

SNR	Average	-10 dB	-5 dB	0 dB	5 dB	10 dB
Unprocessed	-0.04 ± 0.44	-10.02 ± 0.03	-5.03 ± 0.01	-0.03 ± 0.01	4.95 ± 0.01	9.90 ± 0.02
DNN-WF	6.92 ± 0.42	-1.96 ± 0.66	3.43 ± 0.53	7.25 ± 0.42	11.58 ± 0.38	14.25 ± 0.34
VAE	6.72 ± 0.43	-1.92 ± 0.75	2.99 ± 0.59	6.89 ± 0.49	11.43 ± 0.42	14.14 ± 0.37
proposed NA-VAE	**7.29 ± 0.43**	**-1.00 ± 0.78**	**3.64 ± 0.59**	**7.30 ± 0.50**	**11.85 ± 0.42**	**14.57 ± 0.39**

Table 1. Performance comparison in SI-SDR on 5 different SNR conditions trained and evaluated on different subsets of the QUT-NOISE dataset (4 noise types). Values of SI-SDR are given in mean ± confidence interval (95% confidence) over all utterances of the evaluation dataset with unit dB. NA-VAE refers to the proposed noise-aware VAE.

SNR	Average	-10 dB	-5 dB	0 dB	5 dB	10 dB
Unprocessed	-0.04 ± 0.44	-10.01 ± 0.01	-5.02 ± 0.01	-0.03 ± 0.01	4.95 ± 0.01	9.90 ± 0.02
DNN-WF	2.93 ± 0.45	-7.38 ± 0.38	-1.65 ± 0.26	3.25 ± 0.24	8.07 ± 0.22	12.34 ± 0.21
VAE	11.44 ± 0.54	2.74 ± 1.20	7.90 ± 1.07	12.27 ± 0.90	15.27 ± 0.72	19.02 ± 0.68
proposed NA-VAE	**11.88 ± 0.52**	**3.45 ± 1.10**	**8.60 ± 1.03**	**12.70 ± 0.89**	**15.63 ± 0.71**	**19.06 ± 0.67**

Table 2. Performance comparison in SI-SDR on 5 different SNR conditions trained on the QUT-NOISE dataset and evaluated on the DEMAND dataset (12 noise types, completely unseen noise conditions). Values of SI-SDR are given in mean ± confidence interval (95% confidence) over all utterances of the evaluation dataset with unit dB.

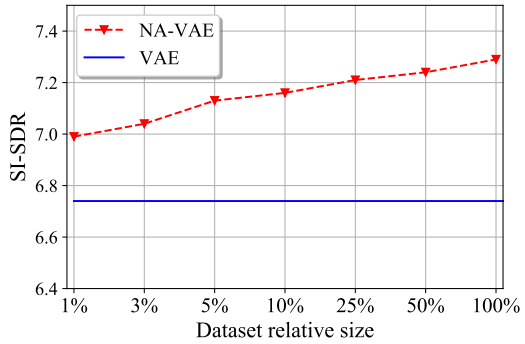


Fig. 3. Influence of the amount of noisy-clean speech training data on SI-SDR improvements for both VAE models, averaged over all noise conditions.

5. RESULTS AND DISCUSSIONS

5.1. Performance evaluation

As can be seen from the results in Table 1 which presents results trained and evaluated on different subsets of QUT-NOISE, the proposed noise-aware VAE outperforms the standard VAE in terms of overall distortion in all SNR scenarios, and the SI-SDR improvements are more evident at low SNR conditions. For example, the noise-aware VAE outperforms the baseline VAE by nearly 1 dB at an input SNR of -10 dB. Table 1 also shows that the DNN-WF performs better than the plain VAE, which implies that appropriate prior noise information is beneficial. In Table 2, which shows the evaluation performed on the DEMAND database while training is still conducted on QUT-NOISE, we see that the fully connected DNN-WF performs significantly worse than the other models. This was expected as we now test on a different more diverse dataset with 12 noise types instead of only 4. The supervised DNN-WF can not transfer the denoising capability to unseen noise types implying that inappropriate prior noise information may even deteriorate performance [8, 14]. However, the proposed noise-aware VAE can still outperform VAE in all SNR conditions, which suggests that the proposed method of improving latent variables in the latent space under this configuration is more capable of generalizing to unseen noise scenarios. Informal listening confirms the SI-SDR results especially for Table 1, while the improvements reported in Table 2 are relatively subtle. Audio examples are available online [1].

5.2. Analysis of the amount of training data

We then look at the influence of the amount of noisy-clean speech training data for estimating the speech latent variable. To achieve this, we initialize the noise-aware encoder with the encoder parameters of the pre-trained standard VAE and then train the new encoder by randomly selecting 1%, 3%, 5%, 10%, 25%, 50% of the noisy-clean speech pairs constructed with the QUT-NOISE dataset. In Fig. 3, it is shown that the performance can already be improved by using only a small percentage of the paired noisy-clean speech data. A value of more than 0.2 dB SI-SDR improvement can be observed with just 1% of the total paired data. It can also be observed that increasing the number of data in the later stage leads to gradual improvements, which may be due to the noise diversity already being largely represented in the small fraction of data used. The research can be extended by increasing the diversity of the noise types in the training phase. This ability of improving performance with only few labeled data shows potential in alleviating overfitting issues in supervised training strategies.

6. CONCLUSION

In this paper, we proposed a noise-aware encoding scheme to improve the robustness of the VAE encoder particularly in low SNRs. For this we incorporate noise information into the VAE encoder to enable a more accurate speech variance estimation based on improved latent variables. By constraining the latent space, the VAE with the proposed noise-aware encoder can learn a non-linear mapping from the noisy mixture to latent variables that represent the clean speech statistics. Our proposed VAE outperforms the standard VAE and a supervised DNN-based filter in SI-SDR. Experiments also showed the generalization ability to unseen noise scenarios by evaluating across different datasets. Moreover, we showed that we could improve the performance even with a small amount of noisy-clean speech data. For future work, our approach could also be integrated with deep generative models that combine temporal dependencies [27].

[1] https://uhh.de/inf-sp-navae2021

7. REFERENCES

[1] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2011.

[2] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*, Morgan & Claypool Publishers, 2013.

[3] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[4] C. Févotte, N. Bertin, and J.L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[5] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear mmse filter for single channel speech enhancement based on non-negative matrix factorization," in *IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*. IEEE, 2011, pp. 45–48.

[6] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[7] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[8] R. Rehr and T. Gerkmann, "An analysis of noise-aware features in combination with the size and diversity of training data for dnn-based speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 601–605.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[10] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations*, Yoshua Bengio and Yann LeCun, Eds., 2014.

[11] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *International Conference on Machine Learning*, p. 1278–1286, 2014.

[12] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in neural information processing systems*, 2014, pp. 3581–3589.

[13] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in neural information processing systems*, 2015, pp. 3483–3491.

[14] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 716–720.

[15] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *International Workshop on Machine Learning for Signal Processing (MLSP)*, 2018, pp. 1–6.

[16] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 101–105.

[17] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, SM. A. Eslami, D. J. Rezende, and O. Ronneberger, "A probabilistic u-net for segmentation of ambiguous images," in *Advances in Neural Information Processing Systems*, 2018, pp. 6965–6975.

[18] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3155–3167, 2011.

[19] G. C. Wei and M. A. Tanner, "A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms," *Journal of the American statistical Association*, vol. 85, no. 411, pp. 699–704, 1990.

[20] C. Robert and G. Casella, *Monte Carlo statistical methods*, Springer Science & Business Media, 2013.

[21] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) sennheiser ldc93s6b," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[22] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The qut-noise-timit corpus for the evaluation of voice activity detection algorithms," *Proceedings Interspeech*, 2010.

[23] J. Thiemann, N. Ito and E. Vincent, "DEMAND: Diverse Environments Multichannel Acoustic Noise Database," http://parole.loria.fr/DEMAND/, 2013.

[24] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 577–581.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[26] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[27] J. Richter, G. Carbajal, and T. Gerkmann, "Speech Enhancement with Stochastic Temporal Convolutional Networks," in *Proceedings Interspeech*, 2020.

3.2 Joint Reduction of Ego-Noise and Environmental Noise with a Partially-Adaptive Dictionary [P7]

Reference

H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Joint reduction of ego-noise and environmental noise with a partially-adaptive dictionary," in *Speech Communication; 14th ITG Conference*, online, 2021, pp. 1–5.

Copyright Notice

Joint Reduction of Ego-noise and Environmental Noise with a Partially-adaptive Dictionary

Huajian Fang[1,2]*, Guillaume Carbajal*[1]*, Stefan Wermter*[2]*, Timo Gerkmann*[1]

[1]Signal Processing (SP), Universität Hamburg, Hamburg, Germany
[2]Knowledge Technology (WTM), Universität Hamburg, Hamburg, Germany
Email: {fang, carbajal, wermter, gerkmann}@informatik.uni-hamburg.de

Abstract

We consider the problem of simultaneous reduction of *ego-noise*, i.e., the noise produced by a robot, and environmental noise. Both noise types may occur simultaneously for humanoid interactive robots. Dictionary- and template-based approaches have been proposed for ego-noise reduction. However, most of them lack adaptability to unseen noise types and thus exhibit limited performance in real-world scenarios with environmental noise. Recently, a variational autoencoder (VAE)-based speech model combined with a fully-adaptive dictionary-based noise model, i.e., non-negative matrix factorization (NMF), has been proposed for environmental noise reduction, showing decent adaptability to unseen noise data. In this paper, we propose to extend this framework with a partially-adaptive dictionary-based noise model, which partly adapts to unseen environmental noise while keeping the part pre-trained on ego-noise unchanged. With appropriate sizes, we demonstrate that the partially-adaptive approach outperforms the approaches based on the fully-adaptive and completely-fixed dictionaries, respectively.

1 Introduction

Interactive robots have attracted a great deal of attention in the last decades. Intelligent interaction between humans and robots relies on robust verbal communication. However, the speech dialogue system of robots is not only affected by environmental noise, but also severely disturbed by *ego-noise*, which refers to the self-created noise mainly generated by electrical and mechanical elements of robots [1].

Ego-noise reduction is a challenging task [2]. The robots' microphones are placed close to the motors, which is mainly the case for small sized robots, resulting in challenging low signal-to-noise ratios (SNRs). Ego-noise is highly nonstationary due to irregular movements of the robot at different speeds, which may degrade the performance of traditional noise tracking algorithms that work independently in each time-frequency bin [3]. However, the spectral structure of ego-noise gives rise to many template- and dictionary-based algorithms [4–8].

Conventional template-based methods estimate ego-noise by selecting pre-learned templates based on command data [9] or motor data [7], but require synchronization of physical state data with audio data. Alternatively, dictionary-based algorithms aim at approximating ego-noise by combining pre-learned feature elements stored in the dictionary. Non-negative matrix factorization (NMF) is a widely used dictionary learning approach [10–12], and it has been shown to be effective in suppressing ego-noise [4, 5]. Incorporating multimodal information, e.g., motor data, into the dictionary-based methods has also been proposed [6, 13]. However, since most dictionary- and template-based algorithms do not consider environmental noise, a fixed template or dictionary trained on ego-noise only can cause noise-mismatch problems in real-world scenarios. To introduce adaptation flexibility, Ince et al. combined a template-based ego-noise estimation algorithm with an independent background noise estimation method [8]. Yet, as their approach only takes stationary noise into account, it results in limited performance in realistic scenarios.

Recently, there has been great interest in deep generative models, such as generative adversarial networks (GANs) [14] and variational autoencoders (VAEs) [15]. In speech enhancement, VAEs have been used to learn a prior distribution of clean speech and have been combined with an untrained NMF noise model to estimate the signal variances using a Monte Carlo expectation maximization (MCEM) algorithm [16–20]. However, this approach assumes only one type of noise, i.e., environmental noise, during inference. As no prior noise information is considered in the model or data, gaining robustness against noise in such a framework remains a challenge [21].

In this work, to overcome the noise mismatch problem and improve noise robustness, we extend the VAE-NMF framework with a partially-adaptive noise dictionary to jointly reduce ego-noise and environmental noise for robots. More specifically, the noise dictionary is split into non-adaptive and adaptive parts, where the non-adaptive part is trained on ego-noise only and fixed during inference, while the untrained adaptive part is used to fit unseen noise, such as environmental noise. We illustrate the benefits of including prior noise information for ego-noise and retaining adaptation flexibility for environmental noise. We show that, with appropriate sizes, the partially-adaptive dictionary approach improves enhancement performance in comparison to the approaches based on the completely-fixed and fully-adaptive noise dictionaries, respectively.

In Section 2, we present the background related to signal modeling and parameter optimization. The proposed partially-adaptive approach is introduced in Section 3, followed by experimental setup in Section 4, results in Section 5, and conclusions in Section 6.

2 Background

2.1 Mixture Model

In the time-frequency domain using the short time Fourier transform (STFT), the mixture signal $x_{ft} \in \mathbb{C}$ is the sum of clean speech $s_{ft} \in \mathbb{C}$ and noise $b_{ft} \in \mathbb{C}$ as:

$$x_{ft} = \sqrt{g_t} s_{ft} + b_{ft}, \tag{1}$$

at the time frame $t \in [1, 2, \cdots, T]$ and the frequency bin $f \in [1, 2, \cdots, F]$, where T denotes the number of time frames and F the number of frequency bins of the utterance. A time-dependent and frequency-independent gain $g_t \in \mathbb{R}_+$ is applied to improve the robustness to the time-varying loudness of different speech sounds [16].

Assuming that all signals are Gaussian variables and uncorrelated, the noisy mixture follows:

$$x_{ft} \sim \mathcal{N}_{\mathbb{C}}(0,\, g_t \nu_{s,ft}^2 + \nu_{b,ft}^2), \tag{2}$$

where $g_t \nu_{s,ft}^2$, $\nu_{b,ft}^2$ represent the variance of speech and noise, and $\mathcal{N}_{\mathbb{C}}$ denotes complex Gaussian distribution. The clean

speech s_{ft} can be estimated by Wiener filtering, which is optimal in the minimum mean square error sense [22]. The Wiener filter can be constructed by:

$$\hat{m}_{ft} = \frac{g_t \hat{\nu}^2_{s,ft}}{g_t \hat{\nu}^2_{s,ft} + \hat{\nu}^2_{b,ft}}, \tag{3}$$

where $g_t \hat{\nu}^2_{s,ft}$, $\hat{\nu}^2_{b,ft}$ represent the estimated variance of speech and noise. Under the local stationarity assumption, the power spectral density of the signal is characterized by the variance [23].

2.2 Noise Model

NMF factorizes a non-negative input matrix into a dictionary matrix and a coefficient matrix; it is widely used to approximate the noise variances [16, 24, 25]. When only the ego-noise $e_{ft} \in \mathbb{C}$ is present, i.e., $b_{ft} = e_{ft}$, the product of the dictionary matrix $\overline{W} \in \mathbb{R}_+^{F \times L}$ and the coefficient matrix $\overline{H} \in \mathbb{R}_+^{L \times T}$ approximates the ego-noise variance, given as:

$$\hat{\nu}^2_{e,ft} = \left(\overline{W}\overline{H}\right)_{ft} = \sum_l \overline{w}_{fl} \overline{h}_{lt}, \tag{4}$$

where L denotes the dictionary size of the ego-noise model, indexed by l [5, 6]. Since ego-noise is characterized by structured patterns in the time-frequency domain, it is advantageous to pre-train and preserve the dictionary matrix [4–6]. When only the environmental noise $n_{ft} \in \mathbb{C}$ is present, i.e., $b_{ft} = n_{ft}$, the environmental noise variance is defined similarly as:

$$\hat{\nu}^2_{n,ft} = (WH)_{ft} = \sum_k w_{fk} h_{kt}, \tag{5}$$

where K denotes the dictionary size of the environmental noise model, indexed by k [16].

2.3 Speech Model

In dictionary-based ego-noise reduction approaches, it is common to model the speech variances using NMFs, e.g., in [4, 5]. Recently, it has been proposed to model speech using a pre-trained VAE, exhibiting better performance than the NMF-based speech model [16].

The VAE provides a scheme that can jointly train generative (decoder) and inference (encoder) models, allowing to infer a latent variable $z_t \in \mathbb{R}^D$ from noisy observation $x_t \in \mathbb{R}^F$ and generate clean speech from the latent variable z_t. Specifically, the VAE is trained on the periodograms of clean speech $|s_t|^2$ [16, 17]. At test time, the clean speech power spectra $\hat{\sigma}^2_s(z_t)$ are generated from latent variables z_t inferred from the observed noisy periodograms $|x_t|^2 \in \mathbb{R}_+^F$. Here, $\hat{\sigma}^2_s : \mathbb{R}^D \to \mathbb{R}_+^F$ denotes the nonlinear mapping from the latent space to the reconstructed speech achieved by the generative model. Although VAE-based speech models have been widely used for environmental noise reduction, they have not been employed in the task of ego-noise reduction.

2.4 Parameter Optimization

In the conventional dictionary-based ego-noise reduction framework [4, 5], where speech and noise are all modeled by NMFs, the dictionary matrix $\overline{W}$ is pre-trained on ego-noise only. At test time, the noise activation matrix $\overline{H}$ and the speech parameters, i.e., the speech dictionary matrix and the speech coefficient matrix, are estimated from the noisy observation using, e.g., multiplicative update rules [24].

Recently, it has been proposed to combine a VAE-based speech model with an untrained NMF-based noise model for environmental noise reduction [16]. The speech model parameters, i.e., the parameters of the VAE model, are obtained by training the neural network on clean speech data, while the noise model parameters, i.e., the noise activation matrix H and the noise coefficients matrix W, are estimated based on the input noisy observation. Since it is intractable to directly compute the maximum likelihood of the model with latent variables z_t and unknown noise parameters, the MCEM algorithm has been alternatively proposed to solve the parameter estimation problem [16, 26, 27].

The conventional ego-noise reduction algorithms based on a completely fixed dictionary may lack the adaptability to unseen noise data, resulting in limited performance in realistic scenarios. The VAE-NMF framework considers only environmental noise. It maintains the flexibility of adaptation, but includes no prior noise information in the model. Therefore, its enhancement performance may be lower than that of algorithms with prior noise information due to noise-sensitivity issues.

3 Proposed Approach

As in many realistic scenarios ego-noise and environmental noise are simultaneously present, we propose to extend the VAE-NMF framework with a partially-adaptive dictionary. The non-adaptive pre-trained part allows for an efficient modeling of ego-noise, while the adaptive untrained part maintains the ability to adapt to unseen environmental noise.

3.1 Mixture Model

Assuming a more realistic scenario where robotic ego-noise e_{ft} and environmental noise n_{ft} are simultaneously present, the noise signal b_{ft} consists of the superposition of e_{ft} and n_{ft}, as:

$$b_{ft} = e_{ft} + n_{ft}. \tag{6}$$

We propose to use NMFs to model these noise signals separately. Given the VAE speech model and two NMF noise models, under the assumption of uncorrelated Gaussian variables, the noisy mixture can be described by:

$$x_{ft} \sim \mathcal{N}_{\mathbb{C}}(0,\ g_t \hat{\sigma}^2_{s,f}(z_t) + \sum_k w_{fk} h_{kt} + \sum_l \overline{w}_{fl} \overline{h}_{lt}), \tag{7}$$

where $\hat{\sigma}^2_{s,f} : \mathbb{R}^D \to \mathbb{R}_+$ denotes the nonlinear function $\hat{\sigma}^2_s$ for the f-th frequency bin. The Wiener filter can be constructed by:

$$\hat{m}_{ft} = \frac{g_t \hat{\sigma}^2_{s,f}(z_t)}{g_t \hat{\sigma}^2_{s,f}(z_t) + \sum_k w_{fk} h_{kt} + \sum_l \overline{w}_{fl} \overline{h}_{lt}}. \tag{8}$$

To find the maximum likelihood solution for the model with the latent variable z_t and the unknown parameters $\zeta = \{w_{fk},\ h_{kt},\ \overline{w}_{fl},\ \overline{h}_{lt},\ g_t\}$, we adapt the MCEM algorithm by Leglaive et al. [16] to the proposed partially-adaptive dictionary approach.

3.2 Parameter Optimization

We describe the E-step and M-step of the MCEM algorithm related to the proposed partially-adaptive dictionary approach.

3.2.1 E-step

In the E-step, we compute the expectation of the complete data log-likelihood. As in Leglaive et al. [16], the intractable posterior distribution is approximated by the Metropolis-Hastings

algorithm. The integral of the expectation is approximated by the average of R samples, indexed by r, as:

$$Q(\zeta,\zeta^{\sharp}) = E_{p(Z|X,\zeta^{\sharp})}[\ln(p(X,Z|\zeta))]$$
$$\simeq -\frac{1}{R}\sum_{r}\left(\sum_{f,t}\left(\ln\left(g_t\hat{\sigma}^2_{s,f}(z_t^{(r)})+\sum_k w_{fk}h_{kt}+\sum_l \overline{w}_{fl}\overline{h}_{lt}\right)\right.\right.$$
$$\left.\left.+\frac{|x_{ft}|^2}{g_t\hat{\sigma}^2_{s,f}(z_t^{(r)})+\sum_k w_{fk}h_{kt}+\sum_l \overline{w}_{fl}\overline{h}_{lt}}\right)+TF\ln(\pi)\right), \tag{9}$$

where the last constant term can be ignored and $\zeta^{\sharp}$ denotes an initialization of the parameters.

3.2.2 M-step

In the M-step, we optimize $Q(\zeta,\zeta^{\sharp})$ with respect to the model parameters $\zeta = \{w_{fk},\, h_{kt},\, \overline{w}_{fl},\, \overline{h}_{lt},\, g_t\}$. Maximum likelihood estimation of the dictionary matrix and the coefficient matrix from input observation is equivalent to NMF optimization using the *Itakura-Saito (IS) divergence* [25]. Thus, we may use the maximization-minimization (MM) algorithm introduced in [24] to estimate the unknown model parameters. We provide for a similar derivation as in [16, 24] in a supporting document for the sake of completeness [1]. The update rules of h_{kt}, $\overline{h}_{lt}$, w_{fk} and g_t are given as:

$$h_{kt} = h_{kt}\cdot\left(\frac{\sum_f w_{fk}\cdot|x_{ft}|^2\cdot\sum_r\left(V^{(r)}_{x,ft}\right)^{-2}}{\sum_f w_{fk}\cdot\sum_r\left(V^{(r)}_{x,ft}\right)^{-1}}\right)^{\frac{1}{2}}, \tag{10}$$

$$\overline{h}_{lt} = \overline{h}_{lt}\cdot\left(\frac{\sum_f \overline{w}_{fl}\cdot|x_{ft}|^2\cdot\sum_r\left(V^{(r)}_{x,ft}\right)^{-2}}{\sum_f \overline{w}_{fl}\cdot\sum_r\left(V^{(r)}_{x,ft}\right)^{-1}}\right)^{\frac{1}{2}}, \tag{11}$$

$$w_{fk} = w_{fk}\cdot\left(\frac{\sum_t |x_{ft}|^2\cdot\sum_r\left(V^{(r)}_{x,ft}\right)^{-2}\cdot h_{kt}}{\sum_t h_{kt}\cdot\sum_r\left(V^{(r)}_{x,ft}\right)^{-1}}\right)^{\frac{1}{2}}, \tag{12}$$

$$g_t = g_t\cdot\left(\frac{\sum_f |x_{ft}|^2\cdot\sum_r V^{(r)}_{s,ft}\cdot\left(V^{(r)}_{x,ft}\right)^{-2}}{\sum_f\sum_r V^{(r)}_{s,ft}\cdot\left(V^{(r)}_{x,ft}\right)^{-1}}\right)^{\frac{1}{2}}, \tag{13}$$

where $V^{(r)}_{x,ft} = g_t\hat{\sigma}^2_{s,f}(z_t^{(r)})+\sum_k w_{fk}h_{kt}+\sum_l \overline{w}_{fl}\overline{h}_{lt}$ and $V^{(r)}_{s,ft} = \hat{\sigma}^2_{s,f}(z_t^{(r)})$. Note that the dictionary of the ego-noise model $\overline{W}$ is fixed during the M-step. We learn the dictionary of the ego-noise model $\overline{W}$ during the training phase which we detail in the next subsection.

3.3 Training the Ego-Noise Dictionary

During the training phase, we train a single NMF model on ego-noise only using the IS divergence to obtain ego-noise features [24], shown as:

$$C_{\text{IS}} = \sum_{f,t}\left(\frac{|e_{ft}|^2}{(\overline{W}\overline{H})_{ft}} - \log\left(\frac{|e_{ft}|^2}{(\overline{W}\overline{H})_{ft}}\right) - 1\right), \tag{14}$$

[1] https://uhh.de/inf-sp-partiallyadaptive2021

where $|e_{ft}|^2$ indicates the periodogram of ego-noise at frequency f and time t. At test time, we set $\overline{W}$ to the dictionary trained solely on ego-noise, while its coefficient matrix $\overline{H}$ is adaptive to the input mixture as described above.

4 Experimental Setup

The proposed algorithm is evaluated with signals we recorded in our varechoic chamber and real environmental noise from the DEMAND dataset [28].

4.1 Dataset

The experimental evaluation is conducted using a humanoid interactive robot NAO H25 from Softbank [29]. We created ego-noise by performing right arm movements involving 6 joints while the robot was in the crouch posture. We recorded the signals by an omnidirectional electret microphone that was mounted externally to the robot close to the position of the built-in front head microphone. In total, we generated approximately 7 min of ego-noise. To create the training data for the VAE-based speech model, we recorded the TIMIT training set [30] by playing back the noise-free speech sentences through a loudspeaker placed at 1 m distance from the robot's front at a height of 1.2 m. The recordings were made in our varechoic chamber with the reverberation time set to $T_{60} = 200$ ms. All time-domain audio signals are sampled at a rate of 16 kHz.

For the evaluation, we create three evaluation datasets, imitating three different human-robot talking scenarios. We first select 840 speech samples from the TIMIT test set re-recorded with the same setup as above.

1. *Ego-noise only*: To simulate the talking scenario where ambient sounds can be ignored, we mix speech signals with out-of-training ego-noise only at an SNR randomly sampled in the range of $[-5, +5]$ dB with a gap of 1 dB.
2. *Env-noise only*: To mimic the talking scenario where the robot stops moving and only environmental sounds are present, we add the environmental noise selected from the DEMAND database to speech samples [28]. The environmental noise signals are randomly sampled from the noise data in the categories {*domestic*, *public*, *street*, *transportation*} and added at a random SNR chosen in the range of [2, 14] dB with a gap of 3 dB.
3. *Ego + Env*: To imitate a realistic scenario where a person is talking to a moving robot while environmental noise is present simultaneously, we corrupt speech samples with both noise types. For each noise type, a random SNR in the same range as above is used.

4.2 Baselines

We compare the VAE-NMF framework based on different adaptive schemes: the completely-fixed dictionary (denoted as *NMF-fixed*), the fully-adaptive dictionary (denoted as *NMF-full*), and the proposed partially-adaptive dictionary (denoted as *NMF-partial*).

4.3 Parameter Settings

Total dictionary size	16	32	64	96	128
K	8	16	32	32	32
L	8	16	32	64	96

Table 1: Dictionary sizes.

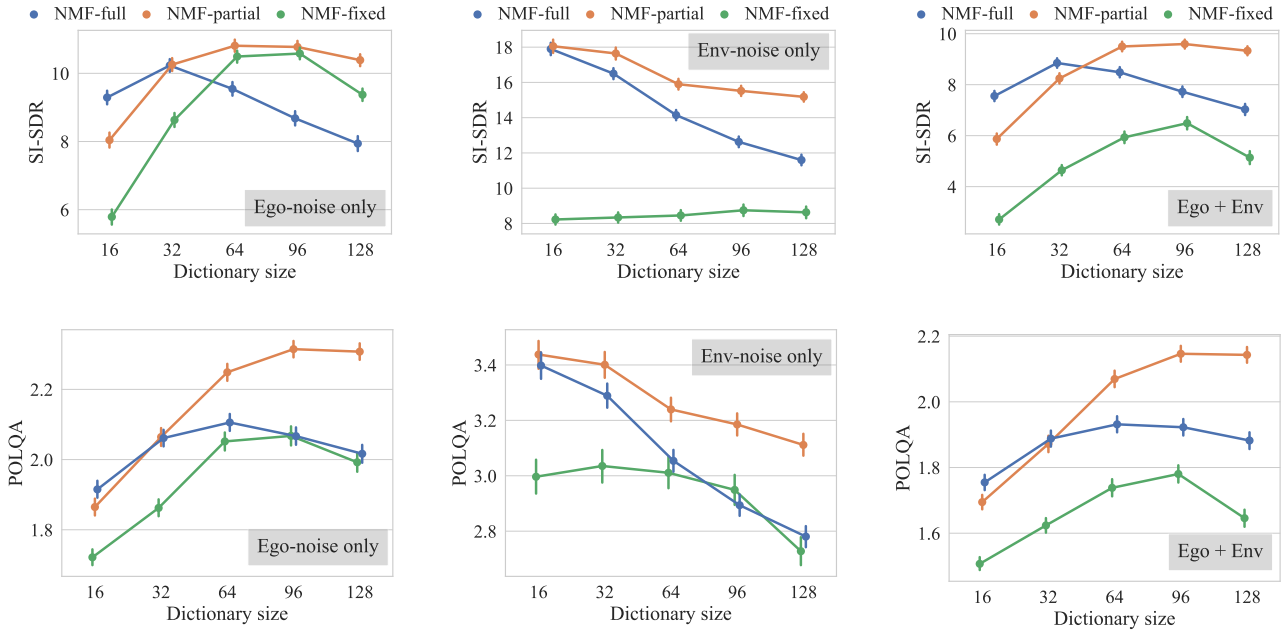


Figure 1: Performance of three adaptive schemes evaluated at different NMF dictionary sizes in SI-SDR (in dB) and POLQA, for three different corrupted scenarios: *Ego-noise only*, *Env-noise only*, and *Ego + Env*. For both metrics, higher numbers indicate better performance. The marker denotes the mean value over all utterances in the test dataset and the vertical bar indicates the 95%-confidence interval.

The signals are transformed into the time-frequency domain using the STFT with a Hann window, where the window length and the hop size are set to 64 ms and 16 ms, respectively. We test the different dictionary sizes when modeling noise signals, as shown in Table 1. For total NMF dictionary sizes greater than 64, $K = 32$ empirically shows better performance than the other tested cases. The composing matrices W, H, and $\overline{H}$ are randomly initialized whereas the dictionary matrix $\overline{W}$ of the ego-noise model is pre-trained. The parameters of the Metropolis-Hastings algorithm follow the setting in [16].

Both the encoder and the decoder of the VAE have two feed-forward hidden layers of 128 units, and the hyperbolic tangent activation function is applied to all hidden layers. The dimension of the latent space is fixed at 16 for all algorithms. The parameters of the VAE are optimized by Adam [31] with a learning rate of 1×10^{-3}.

4.4 Evaluation Metrics

We measure the performance of ego-noise reduction using the scale-invariant signal-to-distortion ratio (SI-SDR) measured in dB [32], which indicates the overall distortion including noise reduction and speech artifacts, and the perceptual objective listening quality analysis (POLQA)[2] [33], which measures speech quality.

5 Results

From left to right, the three columns in Figure 1 show evaluation results for cases corrupted by ego-noise only, by environmental noise only, and by both ego-noise and environmental noise respectively. The enhancement performance in the presence of ego-noise only shows that with proper noise information and dictionary sizes, NMF-partial and NMF-fixed perform similarly in SI-SDR and slightly better than NMF-full. NMF-partial improves speech quality in comparison to NMF-full and NMF-fixed. When prior ego-noise information fails to match the test case as shown in the second column, the proposed NMF-partial still gives competitive or better performance than NMF-full. The performance gain may come from the smaller size of the adaptive part of the dictionary, because the model performance as shown by NMF-full can benefit from a small adaptive dictionary size. The results shown in the first two columns indicate that the proposed NMF-partial with appropriate NMF dictionary sizes does not deteriorate the system even if one of the assumed noise types is missing.

In the third column, where ego-noise and environmental noise are present simultaneously, NMF-partial is superior to NMF-fixed among all dictionary sizes considered, which shows the interest of using the partially-adaptive scheme and not only a fixed dictionary to extract noise components from an unseen noisy mixture. For example, NMF-partial significantly outperforms NMF-fixed by more than 3 dB in SI-SDR at the NMF dictionary size of 96. A performance improvement of around 0.75 dB can also be observed between NMF-partial (at the size of 96) and NMF-full (at the size of 32), indicating that the partially-adaptive scheme rather than adaptively estimating all noise components is favorable. Similarly, we observe that NMF-partial reaches the peak in terms of POLQA at the dictionary size of 96 and outperforms both NMF-full and NFM-fixed. Audio examples and the recorded ego-noise data are available online [1].

6 Conclusions

In this paper, we introduced a partially-adaptive noise dictionary to jointly reduce ego-noise and environmental noise for interactive NAO robots. The proposed partially-adaptive scheme can improve noise robustness by incorporating prior ego-noise information and retains the flexibility to adapt to unseen noise data. With appropriate dictionary sizes, the presented approach showed superior performance over the methods based on the fully-adaptive dictionary and the completely-fixed dictionary in complex situations where ego-noise and environmental noise are present simultaneously.

[2]We would like to thank J. Berger and Rohde&Schwarz SwissQual AG for their support with POLQA.

References

[1] J. Dávila-Chacón, J. Liu, and S. Wermter, "Enhanced robot speech recognition using biomimetic binaural sound source localization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 1, pp. 138–150, 2019.

[2] A. Schmidt, H. W. Löllmann, and W. Kellermann, "Acoustic self-awareness of autonomous systems in a world of sounds," *Proceedings of the IEEE*, vol. 108, no. 7, pp. 1127–1149, 2020.

[3] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[4] T. Tezuka, T. Yoshida, and K. Nakadai, "Ego-motion noise suppression for robots based on semi-blind infinite non-negative matrix factorization," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6293–6298, 2014.

[5] T. Haubner, A. Schmidt, and W. Kellermann, "Multichannel nonnegative matrix factorization for ego-noise suppression," in *Speech Communication; 13th ITG-Symposium*, pp. 1–5, 2018.

[6] A. Schmidt, A. Brendel, T. Haubner, and W. Kellermann, "Motor data-regularized nonnegative matrix factorization for ego-noise suppression," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–15, 2020.

[7] G. Ince, K. Nakadai, T. Rodemann, J. Imura, K. Nakamura, and H. Nakajima, "Incremental learning for ego noise estimation of a robot," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 131–136, 2011.

[8] G. Ince, K. Nakadai, T. Rodemann, J.-i. Imura, K. Nakamura, and H. Nakajima, "Assessment of single-channel ego noise estimation methods," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 106–111, IEEE, 2011.

[9] Y. Nishimura, M. Ishizuka, K. Nakadai, M. Nakano, and H. Tsujino, "Speech recognition for a humanoid with motor noise utilizing missing feature theory," in *2006 6th IEEE-RAS International Conference on Humanoid Robots*, pp. 26–33, 2006.

[10] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[11] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," in *IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, pp. 45–48, IEEE, 2011.

[12] P. Smaragdis, C. Févotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, 2014.

[13] A. Schmidt, H. W. Löllmann, and W. Kellermann, "A novel ego-noise suppression algorithm for acoustic signal enhancement in autonomous systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6583–6587, IEEE, 2018.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[15] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations* (Y. Bengio and Y. LeCun, eds.), 2014.

[16] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2018.

[17] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 716–720, 2018.

[18] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks," in *Interspeech*, pp. 4516–4520, 2020.

[19] G. Carbajal, J. Richter, and T. Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 681–685, IEEE, 2021.

[20] G. Carbajal, J. Richter, and T. Gerkmann, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement," *arXiv preprint arXiv:2105.08970*, 2021.

[21] H. J. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 676–680, 2021.

[22] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement* (E. Vincent, T. Virtanen, and S. Gannot, eds.), pp. 65–85, Wiley, 2018.

[23] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3155–3167, 2011.

[24] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[25] C. Févotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[26] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[27] G. C. Wei and M. A. Tanner, "A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 699–704, 1990.

[28] J. Thiemann, N. Ito and E. Vincent, "DEMAND: Diverse Environments Multichannel Acoustic Noise Database." http://parole.loria.fr/DEMAND/, 2013.

[29] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, "Mechatronic design of nao humanoid," in *2009 IEEE International Conference on Robotics and Automation*, pp. 769–774, IEEE, 2009.

[30] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," 1993.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[32] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR — half-baked or well done?," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019.

[33] ITU-T Rec. P.863, "Perceptual objective listening quality prediction." *International Telecommunication Union*, 2011.

3.3 Partially Adaptive Multichannel Joint Reduction of Ego-Noise and Environmental Noise [P8]

Reference

H. Fang, N. Wittmer, J. Twiefel, S. Wermter, and T. Gerkmann, "Partially adaptive multichannel joint reduction of ego-noise and environmental noise," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096344

Copyright Notice

PARTIALLY ADAPTIVE MULTICHANNEL JOINT REDUCTION OF EGO-NOISE AND ENVIRONMENTAL NOISE

Huajian Fang[1,2]*, Niklas Wittmer*[1]*, Johannes Twiefel*[2,3]*, Stefan Wermter*[2]*, Timo Gerkmann*[1]

[1]Signal Processing (SP), [2]Knowledge Technology (WTM), Universität Hamburg, Germany
[3]exXxa GmbH, Hamburg, Germany

ABSTRACT

Human-robot interaction relies on a noise-robust audio processing module capable of estimating target speech from audio recordings impacted by environmental noise, as well as self-induced noise, so-called ego-noise. While external ambient noise sources vary from environment to environment, ego-noise is mainly caused by the internal motors and joints of a robot. Ego-noise and environmental noise reduction are often decoupled, i.e., ego-noise reduction is performed without considering environmental noise. Recently, a variational autoencoder (VAE)-based speech model has been combined with a fully adaptive non-negative matrix factorization (NMF) noise model to recover clean speech under different environmental noise disturbances. However, its enhancement performance is limited in adverse acoustic scenarios involving, e.g. ego-noise. In this paper, we propose a multichannel partially adaptive scheme to jointly model ego-noise and environmental noise utilizing the VAE-NMF framework, where we take advantage of spatially and spectrally structured characteristics of ego-noise by pre-training the ego-noise model, while retaining the ability to adapt to unknown environmental noise. Experimental results show that our proposed approach outperforms the methods based on a completely fixed scheme and a fully adaptive scheme when ego-noise and environmental noise are present simultaneously.

***Index Terms*—** Ego-noise reduction, speech enhancement, variational autoencoder, multichannel non-negative matrix factorization

1. INTRODUCTION

In recent decades, research on autonomous systems (AS) such as humanoid interactive robots has received increasing attention. Interactive robots are typically equipped with multiple microphones to perceive their environment and react to requests or particular commands from humans. However, the acquisition of target acoustic information is often disturbed not only by external interfering sources, i.e., environmental noise, but also by self-generated noise, also called *ego-noise*. It poses difficulties for subsequent tasks, such as speech recognition and language understanding. This calls for a noise-robust audio processing module capable of recovering target clean speech to support the robot's actuator unit to act appropriately [1, 2].

In human-robot interaction, ego-noise may originate from different parts of the robot and reducing ego-noise is non-trivial in various aspects. It is mainly caused by the electric motors and mechanical parts distributed all over the robot body [3, 4]. The microphones are often placed close to the motors, especially for small-sized robots, resulting in acoustic scenarios with challenging signal-to-noise ratios (SNRs). Furthermore, as ego-noise coming from, e.g. robotic limb movements, is non-stationary, it may be considered a difficult noise source. However, due to the limited degree of motion, ego-noise from the motors and joints exhibits a characteristic spatial and spectral structure. Thus, specialized and efficient light-weight machine learning algorithms can be designed to learn and exploit these distinct spatial and spectral characteristics of ego-noise [1, 3–10].

This work was funded by the DFG project number 261402652 and ahoi.digital.

For instance, ego-noise can be modeled by dictionary-based algorithms, e.g. non-negative matrix factorization (NMF) [9, 11, 12], where ego-noise is approximated by a linear combination of pre-captured dictionary components. For multichannel recordings, in addition to structured tempo-spectral characteristics, spatial information can also be employed using, e.g. multichannel NMF [6, 7]. Deleforge et al. [4] have proposed a sparse representation of multichannel ego-noise signals in the complex domain. Some approaches have included information from other modalities, such as motor data [5, 8, 12]. However, this requires synchronized multimodal data, which may not be readily available. While a pre-learned ego-noise model has shown some effectiveness in modeling noise characteristics, it may cause noise mismatch problems in realistic scenarios that include not only ego-noise, but also unknown environmental noise signals.

Currently, advanced methods for environmental noise reduction are based on deep neural networks (DNNs) [13]. The variational autoencoder (VAE) is a deep generative model that can be used to learn a probabilistic prior distribution of clean speech [14]. It has been combined with a statistical NMF noise model to perform speech enhancement, where the VAE-based speech model is pre-trained on clean speech while the parameters of the NMF model are estimated based on noisy observations [15–17]. The VAE-NMF framework has shown improved speech enhancement performance and generalization capabilities over its NMF counterpart and fully supervised baselines [15–17]. While the fully adaptive NMF noise model can potentially adapt to various acoustic scenarios, gaining robustness under adverse acoustic conditions (e.g. when ego-noise and environmental noise are present simultaneously) remains a challenging task, as we will show in experiments. Few existing publications take both ego-noise and environmental noise into account [10, 18, 19]. Ince et al. proposed to reduce stationary background noise independently of ego-noise [10]. Our previous work [18] has presented a single-channel joint noise reduction system for interactive robots, but disregarded spatial information.

In this work, we propose a multichannel joint ego-noise and environmental noise reduction method for interactive robots. For this, the tempo-spectral features of speech are modeled using the VAE and the noise characteristics are modeled by multichannel NMF as in [17]. More specifically, similar to multichannel ego-noise approaches such as [7], we want to take advantage of spatially and spectrally structured characteristics of ego-noise to gain robustness in adverse conditions. At the same time, similar to, e.g. [17], we want to retain the adaptation ability to unknown environmental noise. For this, we propose to model ego-noise and environmental noise separately. We pre-train the ego-noise model to capture the spectral and spatial features, while its temporal activation is adapted to noisy observations jointly with the parameters of the environmental noise model. Experimental results show the considerable benefits of the proposed joint reduction method when ego-noise and environmental noise are present simultaneously.

2. SIGNAL MODEL

We consider an acoustic scenario where the target speech signal is disturbed by additive noise and recorded by a microphone array with M channels.

We transform the noisy mixture into the time-frequency domain using the short-time Fourier transform (STFT):

$$\mathbf{X}_{ft} = \sqrt{g_t}\mathbf{S}_{ft} + \mathbf{N}_{ft}, \tag{1}$$

where $\mathbf{X}_{ft} \in \mathbb{C}^M$, $\mathbf{S}_{ft} \in \mathbb{C}^M$, and $\mathbf{N}_{ft} \in \mathbb{C}^M$ represent the complex coefficients of the mixture signal, the speech signal, and the noise signal at the frequency bin $f \in \{1,\cdots,F\}$ and the time frame $t \in \{1,\cdots,T\}$. g_t is a gain parameter to increase the robustness to the time-varying loudness of speech sounds [17]. Note that the noise signals $\mathbf{N}_{ft}$ may contain either ego-noise or environmental noise or both. We aim to recover clean speech with improved quality and intelligibility given only noisy mixtures.

2.1. Noise model

The noise coefficients are assumed to follow a complex Gaussian distribution with zero mean

$$\mathbf{N}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{\Sigma}_{N,ft}), \tag{2}$$

where $\mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}, \mathbf{\Sigma})$ denotes the complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. The covariance matrix is defined as

$$\mathbf{\Sigma}_{N,ft} = \mathbf{R}_{N,f}\sigma^2_{N,ft}, \tag{3}$$

where $\mathbf{R}_{N,f} \in \mathbb{C}^{M\times M}$ is a spatial covariance matrix characterizing the sound propagation process from sources to microphones. $\sigma^2_{N,ft}$ represents the noise spectral variance, which can be modeled using the NMF,

$$\sigma^2_{N,ft} = [\mathbf{W}_N\mathbf{H}_N]_{ft} = \sum_{k=1}^{K} w_{fk}h_{kt}, \tag{4}$$

where $\mathbf{W}_N \in \mathbb{R}_+^{F\times K}$ denotes the dictionary matrix that captures the time-frequency characteristics of noise and $\mathbf{H}_N \in \mathbb{R}_+^{K\times T}$ denotes the coefficient matrix that represents the temporal activity. The noise dictionary contains K atoms indexed by k (K is also referred to here as the dictionary size). We will decompose the noise signal $\mathbf{N}_{ft} = \mathbf{E}_{ft} + \mathbf{B}_{ft}$ into ego-noise $\mathbf{E}_{ft}$ and environmental noise $\mathbf{B}_{ft}$ in Section 3.

2.2. Speech model

We assume that the clean speech coefficients are complex Gaussian-distributed:

$$\mathbf{S}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{\Sigma}_{S,f}(\mathbf{z}_t)), \tag{5}$$

where $\mathbf{\Sigma}_{S,f}(\mathbf{z}_t) = \mathbf{R}_{S,f}\sigma^2_{S,f}(\mathbf{z}_t)$. $\mathbf{R}_{S,f} \in \mathbb{C}^{M\times M}$ is the speech spatial covariance matrix. It is assumed that the speech tempo-spectral power can be inferred from the latent variable $\mathbf{z}_t \in \mathbb{R}^L$, denoted as $\sigma^2_{S,f}(\mathbf{z}_t)$, which can be realized by the generative model of the VAE, i.e., the decoder. Let $\mathbf{s}_t \in \mathbb{C}^F$ be a vector of single-channel clean speech spectra at the t-th time frame. The posterior of the latent variable $q(\mathbf{z}_t|\mathbf{s}_t)$ is approximated by a real-valued Gaussian distribution

$$\mathbf{z}_t|\mathbf{s}_t \sim \mathcal{N}(\mu_z(|\mathbf{s}_t|^2), \sigma_z(|\mathbf{s}_t|^2)), \tag{6}$$

where $\mu_z(|\mathbf{s}_t|^2) : \mathbb{R}_+^F \to \mathbb{R}^L$ and $\sigma_z(|\mathbf{s}_t|^2) : \mathbb{R}_+^F \to \mathbb{R}_+^L$ denote the nonlinear mapping from the power spectrogram to the mean and variance of the latent variable, implemented by the encoder of the VAE, also called the recognition model. The parameters of the VAE can be jointly learned by maximizing the variational lower bound of the log-likelihood $\log p(\mathbf{s}_t)$

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q(\mathbf{z}_t|\mathbf{s}_t)}[\log p(\mathbf{s}_t|\mathbf{z}_t)] - \mathbb{KL}(q(\mathbf{z}_t|\mathbf{s}_t)||p(\mathbf{z}_t)), \tag{7}$$

where $\mathbb{KL}(\cdot|\cdot)$ denotes the Kullback-Leibler divergence and $p(\mathbf{z}_t)$ represents the standard Gaussian prior of $\mathbf{z}_t$.

2.3. Clean speech estimation

With the assumption that the speech and noise signals are independent, the noisy mixture is given by

$$\mathbf{X}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, g_t\mathbf{\Sigma}_{S,f}(\mathbf{z}_t) + \mathbf{\Sigma}_{N,ft}). \tag{8}$$

The parameters of the VAE-based speech model are obtained by training the neural network on clean speech data. At testing, a Monte Carlo expectation maximization (MCEM) method can be employed to estimate the unknown parameters $\{\mathbf{W}_N, \mathbf{H}_N, \mathbf{R}_{N,f}, \mathbf{R}_{S,f}, g_t\}$ [17]. Finally, the multichannel Wiener filter is employed to extract clean speech

$$\widehat{\mathbf{S}}_{ft} = g_t\mathbf{\Sigma}_{S,f}(\mathbf{z}_t)(g_t\mathbf{\Sigma}_{S,f}(\mathbf{z}_t) + \mathbf{\Sigma}_{N,ft})^{-1}\mathbf{X}_{ft}. \tag{9}$$

The fully adaptive scheme in [17] that optimizes the unknown parameters based on noisy inputs, will adapt flexibly to different types of noise without the need of prior information on potential noise structures. The main idea of this approach is to achieve a high generalization ability and robustness to unexpected noise types. However, if accurate prior knowledge is available, it can be very helpful to improve robustness, especially in acoustically challenging environments. Therefore, as ego-noise exhibits a very distinct spatial-spectral structure, prior knowledge can be efficiently exploited by pre-learning the dictionary matrix and the spatial covariance matrix on ego-noise recordings only. However, when only pre-learned on ego-noise, the flexibility and generalization to unseen scenarios is lost. For instance, rather poor performance is to be expected in environmental noise, which limits its applicability in realistic scenarios that contain both environmental noise and background noise.

3. JOINT REDUCTION OF EGO-NOISE AND ENVIRONMENTAL NOISE

In this section, we present a multichannel partially adaptive scheme, where we improve noise modeling capabilities by decomposing noise into ego-noise and environmental noise. This allows us to obtain a robust prior pre-learned on the distinct spatial and spectral characteristics of ego-noise, while retaining the flexibility to adapt to environmental noise signals.

3.1. Mixture model and speech estimation

In a real-world human-robot interaction scenario, a target speech signal may be distorted by ego-noise and environmental noise simultaneously. We, thus, consider a noise model that is comprised of ego-noise $\mathbf{E}_{ft}$ and environmental noise $\mathbf{B}_{ft}$ as follows:

$$\mathbf{N}_{ft} = \mathbf{E}_{ft} + \mathbf{B}_{ft}. \tag{10}$$

By assuming that the ego-noise, environmental noise and speech signals are independent and complex Gaussian distributed, the noisy mixture follows a complex Gaussian of the form:

$$\mathbf{X}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, g_t\mathbf{\Sigma}_{S,f}(\mathbf{z}_t) + \mathbf{\Sigma}_{E,ft} + \mathbf{\Sigma}_{B,ft}), \tag{11}$$

where the covariance matrix of environmental noise is defined as $\mathbf{\Sigma}_{B,ft} = \mathbf{R}_{B,f}[\mathbf{W}_B\mathbf{H}_B]_{ft}$ with $[\mathbf{W}_B\mathbf{H}_B]_{ft} = \sum_{k_b=1}^{K_B} w_{fk_b}h_{k_bt}$, and the covariance matrix of ego-noise as $\mathbf{\Sigma}_{E,ft} = \mathbf{R}_{E,f}[\mathbf{W}_E\mathbf{H}_E]_{ft}$ with $[\mathbf{W}_E\mathbf{H}_E]_{ft} = \sum_{k_e=1}^{K_E} w_{fk_e}h_{k_et}$. K_B and K_E are the sizes of the environmental noise dictionary and the ego-noise dictionary, respectively.

Similarly, clean speech can be estimated by applying the multichannel Wiener filter

$$\widehat{\mathbf{S}}_{ft} = g_t\mathbf{\Sigma}_{S,f}(\mathbf{z}_t)(\mathbf{\Sigma}_{X,ft}(\mathbf{z}_t))^{-1}\mathbf{X}_{ft}, \tag{12}$$

where $\mathbf{\Sigma}_{X,ft}(\mathbf{z}_t) = g_t\mathbf{\Sigma}_{S,f}(\mathbf{z}_t) + \mathbf{\Sigma}_{E,ft} + \mathbf{\Sigma}_{B,ft}$. This requires estimating the unknown parameters $\{\mathbf{W}_E, \mathbf{H}_E, \mathbf{W}_B, \mathbf{H}_B, \mathbf{R}_{S,f}, \mathbf{R}_{E,f}, \mathbf{R}_{B,f},$

$g_t\}$. The following subsections describe the estimation of the ego-noise dictionary matrix $\mathbf{W}_E$ and the spatial covariance matrix $\mathbf{R}_{E,f}$ using the pre-training technique, and an MCEM optimization method to the proposed partially adaptive scheme.

3.2. Training phase

To capture the spectral and spatial characteristics of ego-noise, we train a multichannel NMF model on ego-noise recordings by optimizing the negative log-likelihood:

$$\mathcal{L} = \sum_{f=1,t=1}^{F,T} \mathrm{tr}\Big(\mathbf{E}_{ft}\mathbf{E}_{ft}^H\boldsymbol{\Sigma}_{E,ft}^{-1}\Big) + \ln\det(\boldsymbol{\Sigma}_{E,ft}), \tag{13}$$

where constant terms are omitted [20]. $\mathrm{tr}(\cdot)$ denotes the trace operator; $\det(\cdot)$ denotes the determinant of a matrix; $\cdot^H$ denotes the conjugate transpose. Minimizing this function using the majorization scheme leads to the multiplicative update rules for $\{\mathbf{W}_E, \mathbf{H}_E, \mathbf{R}_{E,f}\}$ [17, 20]. We fix the dictionary matrix $\mathbf{W}_E$ and the spatial covariance matrix $\mathbf{R}_{E,f}$ at the testing phase, while keeping $\mathbf{H}_E$ adaptive to noisy observations to account for different temporal variations.

3.3. Parameter optimization

To estimate the unknown parameters in the testing phase, we follow the MCEM optimization scheme by [17]. At the *Expectation step*, the complete-data log-likelihood is approximated by averaging over R samples:

$$\begin{aligned}\mathbf{Q}(\theta;\theta^*) &= \mathbb{E}_{p(\mathbf{z}|\mathbf{X};\theta)}[\ln p(\mathbf{X},\mathbf{z};\theta)] \\ &\approx -\frac{1}{R}\sum_{r=1}^{R}\sum_{f=1,t=1}^{F,T}\Bigg[\mathrm{tr}\Big(\mathbf{X}_{ft}\mathbf{X}_{ft}^H\Big[\boldsymbol{\Sigma}_{X,ft}\Big(\mathbf{z}_t^{(r)}\Big)\Big]^{-1}\Big) \\ &\quad + \ln\det\Big(\boldsymbol{\Sigma}_{X,ft}\Big(\mathbf{z}_t^{(r)}\Big)\Big)\Bigg].\end{aligned} \tag{14}$$

R samples of the latent variable are drawn using the Metropolis-Hastings algorithm with a Gaussian as a symmetric proposal distribution. θ^* is an initialization of the parameters. At the *Maximization step*, we minimize the loss function, i.e., the negative log-likelihood $-R\mathbf{Q}(\theta;\theta^*)$, with respect to the unknown parameters $\theta = \{\mathbf{H}_E, \mathbf{W}_B, \mathbf{H}_B, \mathbf{R}_{S,f}, \mathbf{R}_{B,f}, g_t\}$ using the auxiliary function technique. For this, equation (14) can be viewed as the superposition of a convex function (the first term) and a concave function (the second term), where the former can be bounded using the Jensen's trace inequality and the latter can be bounded using a first-order Taylor expansion [17, Appendix A]. This gives an upper bound function and computing the partial derivative with respect to each parameter separately leads to the iterative update rules:

$$g_t = g_t^*\left[\frac{\sum_{r=1}^{R}\sum_{f=1}^{F}\sigma_f^2\Big(\mathbf{z}_t^{(r)}\Big)\mathrm{tr}\Big[\mathbf{M}_{ft}^{(r)}\mathbf{R}_{S,f}\Big]}{\sum_{r=1}^{R}\sum_{f=1}^{F}\sigma_f^2\Big(\mathbf{z}_t^{(r)}\Big)\mathrm{tr}\Big[\Big(\boldsymbol{\Sigma}_{X,ft}\Big(\mathbf{z}_t^{(r)}\Big)\Big)^{-1}\mathbf{R}_{S,f}\Big]}\right]^{\frac{1}{2}}, \tag{15}$$

$$w_{fk_b} = w_{fk_b}^*\left[\frac{\sum_{r=1}^{R}\sum_{t=1}^{T}h_{k_bt}\mathrm{tr}\Big[\mathbf{M}_{ft}^{(r)}\mathbf{R}_{B,f}\Big]}{\sum_{r=1}^{R}\sum_{t=1}^{T}h_{k_bt}\mathrm{tr}\Big[\Big(\boldsymbol{\Sigma}_{X,ft}\Big(\mathbf{z}_t^{(r)}\Big)\Big)^{-1}\mathbf{R}_{B,f}\Big]}\right]^{\frac{1}{2}}, \tag{16}$$

$$h_{k_bt} = h_{k_bt}^*\left[\frac{\sum_{r=1}^{R}\sum_{f=1}^{F}w_{fk_b}\mathrm{tr}\Big[\mathbf{M}_{ft}^{(r)}\mathbf{R}_{B,f}\Big]}{\sum_{r=1}^{R}\sum_{f=1}^{F}w_{fk_b}\mathrm{tr}\Big[\Big(\boldsymbol{\Sigma}_{X,ft}\Big(\mathbf{z}_t^{(r)}\Big)\Big)^{-1}\mathbf{R}_{B,f}\Big]}\right]^{\frac{1}{2}}, \tag{17}$$

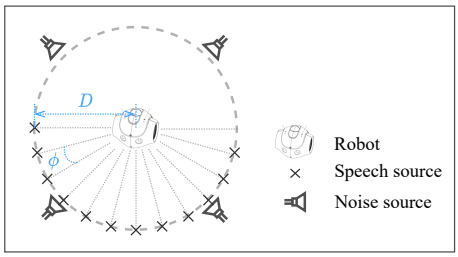


Fig. 1. Illustration of the recording setup with the NAO robot [21]. Room dimensions (length × width × height): 504 × 930 × 284 cm; $T_{60} \approx$ 200 ms; speaker-robot distance $D \approx 1$ m; $\phi \approx 15°$.

$$h_{k_et} = h_{k_et}^*\left[\frac{\sum_{r=1}^{R}\sum_{f=1}^{F}w_{fk_e}\mathrm{tr}\Big[\mathbf{M}_{ft}^{(r)}\mathbf{R}_{E,f}\Big]}{\sum_{r=1}^{R}\sum_{f=1}^{F}w_{fk_e}\mathrm{tr}\Big[\Big(\boldsymbol{\Sigma}_{X,ft}\Big(\mathbf{z}_t^{(r)}\Big)\Big)^{-1}\mathbf{R}_{E,f}\Big]}\right]^{\frac{1}{2}}, \tag{18}$$

where $\mathbf{M}_{ft}^{(r)} = \Big(\boldsymbol{\Sigma}_{X,ft}\Big(\mathbf{z}_t^{(r)}\Big)\Big)^{-1}\mathbf{X}_{ft}\mathbf{X}_{ft}^H\Big(\boldsymbol{\Sigma}_{X,ft}\Big(\mathbf{z}_t^{(r)}\Big)\Big)^{-1}$.

The two adaptive spatial covariance matrices $\mathbf{R}_{S,f}$, $\mathbf{R}_{B,f}$ are updated by solving the corresponding algebraic Riccati equations as in the fully adaptive scheme [17] [20, Appendix I].

4. EXPERIMENTS

In this section, the proposed partially adaptive scheme (referred to as *Partial*) is compared to two baselines:

- *Adaptive*: Refers to the fully adaptive scheme with all unknown parameters estimated based on noisy observations [17].
- *Fixed*: Refers to the fixed scheme with the dictionary matrix and spatial covariance matrix pre-learned on ego-noise recordings at training time and fixed at test time as in the ego-noise reduction literature, e.g. [6, 7].

For each adaptive scheme, 7 different dictionary sizes are considered, leading to a total of 21 compared methods. We evaluate the algorithms in two application scenarios:

- *Ego*: Only ego-noise is present, mimicking a scene where a person is talking to a robot performing certain movements.
- *Ego + Env*: In addition to ego-noise, environmental noise is present simultaneously as an additional disturbance.

We use the scale-invariant signal-to-distortion ratio (Si-SDR) measured in dB to account for both noise reduction and the speech artifacts [22], and the perceptual objective listening quality analysis (POLQA) to measure speech quality [23]. The speech recognition accuracy is measured by the word error rate (WER). We employ the pre-trained speech recognition model Quartznet [24] in the NeMo toolkit [25], in conjunction with a 4-gram language model available via the LibriSpeech website [26].

4.1. Dataset

All algorithms are trained and evaluated on a dataset recorded in our varechoic chamber. We use a humanoid interactive robot NAO H25 from Softbank for recording purposes [27]. The clean speech utterances are randomly chosen from the TIMIT test set [28]. Each target clean speech sample is played through a loudspeaker randomly placed among the positions shown in Fig. 1 and recorded using external omnidirectional electret microphones mounted in the same position of the built-in microphone array ($M = 4$) on the robot. Ego-noise is recorded when the robot performs pre-defined right-arm movements in a crouching posture. To simulate external environmental noise sources, we re-record audio samples randomly selected from the DEMAND database [29] and the loudspeaker emitting environmental noise is placed at one of the four positions shown in Fig. 1. For the ego-noise only

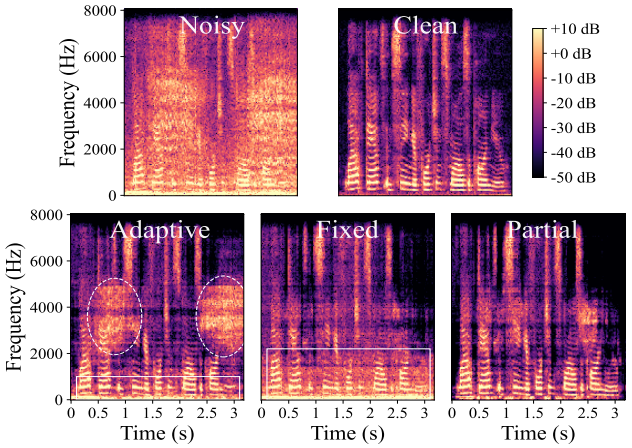


Fig. 2. Spectrograms of an audio example. Clean speech is distorted by both ego-noise and environmental noise. The three plots in the second row represent the reconstructed speech spectrograms obtained by three compared methods.

scenario, we mix speech signals with out-of-training ego-noise recordings (with movement speeds different from training data) at SNRs randomly chosen from{-5 dB, -4 dB, ···, 5 dB}. To simulate a challenging realistic scenario, besides ego-noise, we further corrupt speech signals by environmental noise at a SNR of 0 dB [17]. In total, this leads to a test set of 128 noisy samples for each evaluation scenario, with an average SNR of -2.1 dB for the joint noise scenario and -1.8 dB for the ego-noise only scenario.

4.2. Hyperparameter settings

We use an STFT with a Hann window of 64 ms and a hop size of 25 %. All audio signals are sampled at 16 kHz. The decoder of the VAE has two hidden layers of sizes 128 and 512 respectively. The hyperbolic tangent activation function is applied to the hidden layers; the linear activation is applied to the output layer. The encoder network consists of two hidden layers of sizes 512 and 128, respectively, with the hyperbolic tangent activation functions applied. The latent dimension L is set to 16. The VAE is trained on the re-recorded TIMIT training set using the same microphone setup as described in Section 4.1. The network parameters are optimized using the Adam optimizer with a learning rate of 0.001 and a patience of 5 epochs. The parameters of the MCEM algorithm are set as in [17], i.e., $R=10$ with a burn-in phase 30 iterations. For the partially adaptive scheme, we set the dictionary sizes for the fixed and adaptive parts as shown in Table 1.

Total dictionary size	16	32	64	96	128	160	192
K_B	8	16	32	32	32	32	32
K_E	8	16	32	64	96	128	160

Table 1. Dictionary sizes for the proposed partially adaptive scheme.

4.3. Results

The benefits of the partially adaptive scheme are visible in Fig. 2. While the fully adaptive scheme possesses the flexibility to adapt to various noisy conditions, its ability in capturing noise characteristics is limited especially when both ego-noise and environmental noise are present. This is shown by the residual ego-noise marked with the dashed ellipses and residual environmental noise marked with the solid rectangle in the reconstructed speech spectrogram. While the fixed scheme, whose reconstructed spectrogram is visualized in the second plot in the second row of Fig. 2, shows some effectiveness in removing ego-noise, the residual environmental noise is still quite pronounced, as marked by the solid rectangle. Finally, it can be observed that the proposed partially adaptive scheme shows a higher noise reduction effect than the other two approaches.

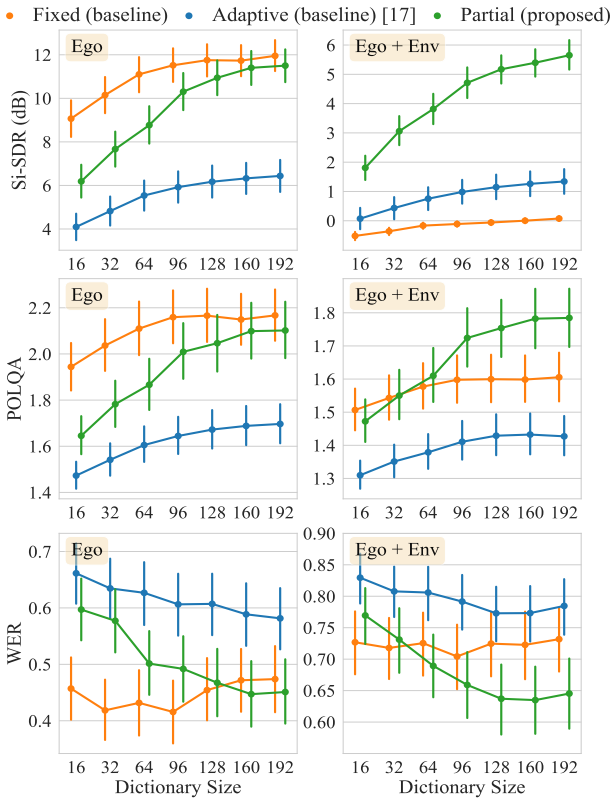


Fig. 3. Higher Si-SDR and POLQA scores indicate better enhancement performance, and lower WER indicates higher recognition accuracy. The marker denotes the mean value and the vertical bar indicates the 95%-confidence interval.

The two columns in Fig. 3 display the evaluation results for the ego-noise only scenario (*Ego*) and the joint noise scenario (*Ego+Env*), respectively. We observe that the fully adaptive approach is outperformed by the fully fixed and partially adaptive schemes in the presence of ego-noise only, as shown by its lowest POLQA and Si-SDR scores and the highest WER. This again implies that the fully adaptive scheme has difficulty in capturing ego-noise characteristics. The partially adaptive scheme and the fully adaptive scheme perform comparably when we increase the total dictionary size, indicating that ego-noise can be better modeled with a larger dictionary size due to its complexity and broadband characteristics. Eventually, it can be observed that the partially adaptive scheme delivers superior results over the other two methods when both noise types are present simultaneously. This indicates that with an appropriate dictionary size, the partially adaptive scheme can effectively approximate ego-noise while properly capturing unknown environmental noise in adverse scenarios. Audio examples are available online[1].

5. CONCLUSION

Based on the deep generative model and multichannel NMF, we proposed to jointly model ego-noise and environmental noise with a partially adaptive scheme. To exploit the spectrally and spatially structured characteristics of ego-noise, we pre-train the ego-noise model while keeping the environmental noise model adaptive to noisy observations. The proposed partially adaptive scheme demonstrated an increased performance compared to the approaches based on the fixed scheme and on the fully adaptive scheme in adverse scenarios where both ego-noise and environmental noise are present.

[1] https://uhh.de/inf-sp-mcpartial2023

6. REFERENCES

[1] Alexander Schmidt, Heinrich W Löllmann, and Walter Kellermann, "Acoustic self-awareness of autonomous systems in a world of sounds," *Proceedings of the IEEE*, vol. 108, no. 7, pp. 1127–1149, 2020.

[2] Jorge Dávila-Chacón, Jindong Liu, and Stefan Wermter, "Enhanced robot speech recognition using biomimetic binaural sound source localization," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 30, no. 1, pp. 138–150, 2019.

[3] Antoine Deleforge, Alexander Schmidt, and Walter Kellermann, "Audio-motor integration for robot audition," in *Multimodal Behavior Analysis in the Wild*, pp. 27–51. Elsevier, 2019.

[4] Antoine Deleforge and Walter Kellermann, "Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Apr. 2015, pp. 355–359.

[5] Alexander Schmidt, Heinrich W. Löllmann, and Walter Kellermann, "A novel ego-noise suppression algorithm for acoustic signal enhancement in autonomous systems," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Apr. 2018, pp. 6583–6587.

[6] Alexander Schmidt and Walter Kellermann, "Multichannel nonnegative matrix factorization with motor data-regularized activations for robust ego-noise suppression," in *IEEE Int. Conf. on Autonomous Systems (ICAS)*, Aug. 2021, pp. 1–5.

[7] Thomas Haubner, Alexander Schmidt, and Walter Kellermann, "Multichannel nonnegative matrix factorization for ego-noise suppression," in *Speech Communication; 13th ITG-Symposium*, Oct. 2018, pp. 1–5.

[8] Akinori Ito, Takashi Kanayama, Motoyuki Suzuki, and Shozo Makino, "Internal noise suppression for speech recognition by small robots," in *Ninth European Conf. on Speech Communication and Technology*, 2005.

[9] Taiki Tezuka, Takami Yoshida, and Kazuhiro Nakadai, "Ego-motion noise suppression for robots based on semi-blind infinite non-negative matrix factorization," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, May 2014, pp. 6293–6298.

[10] Gökhan Ince, Kazuhiro Nakadai, Tobias Rodemann, Jun-ichi Imura, Keisuke Nakamura, and Hirofumi Nakajima, "Assessment of single-channel ego noise estimation methods," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Sept. 2011, pp. 106–111.

[11] Daniel Lee and H Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Proc. Systems*, 2000, vol. 13.

[12] Alexander Schmidt, Andreas Brendel, Thomas Haubner, and Walter Kellermann, "Motor data-regularized nonnegative matrix factorization for ego-noise suppression," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–15, Dec. 2020.

[13] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[14] Diederik P Kingma, Max Welling, et al., "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[15] Simon Leglaive, Laurent Girin, and Radu Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Int. Workshop on Machine Learning for Signal Processing (MLSP)*, Sept. 2018, pp. 1–6.

[16] Kouhei Sekiguchi, Yoshiaki Bando, Kazuyoshi Yoshii, and Tatsuya Kawahara, "Bayesian multichannel speech enhancement with a deep speech prior," in *Asia-Pacific Signal and Information Proc. Association Annual Summit and Conf. (APSIPA ASC)*, Nov. 2018, pp. 1233–1239.

[17] Simon Leglaive, Laurent Girin, and Radu Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2019, pp. 101–105.

[18] Huajian Fang, Guillaume Carbajal, Stefan Wermter, and Timo Gerkmann, "Joint reduction of ego-noise and environmental noise with a partially-adaptive dictionary," in *Speech Communication; 14th ITG Conf.*, Sept. 2021, pp. 1–5.

[19] Gökhan Ince, *Ego Noise Estimation for Robot Audition*, Ph.D. thesis, 2011.

[20] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 21, no. 5, pp. 971–982, 2013.

[21] Aldebaran Robotics, "NAOqi documentation center," `http://doc.aldebaran.com/`.

[22] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "SDR–half-baked or well done?," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2019, pp. 626–630.

[23] ITU-T Rec. P.863, "Perceptual objective listening quality prediction," *Int. Telecommunication Union*, 2011.

[24] Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2020, pp. 6124–6128.

[25] Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al., "Nemo: a toolkit for building ai applications using neural modules," *arXiv preprint arXiv:1909.09577*, 2019.

[26] "openslr.org," `http://openslr.org/11/`, 2022.

[27] David Gouaillier, Vincent Hugel, Pierre Blazevic, Chris Kilner, Jérôme Monceaux, Pascal Lafourcade, Brice Marnier, Julien Serre, and Bruno Maisonnier, "Mechatronic design of NAO humanoid," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, May 2009, pp. 769–774.

[28] John S., Garofolo and Lori F., Lamel and William M., Fisher and Jonathan G., Fiscus and David S., Pallett and Nancy L., Dahlgren and Victor Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic data consortium, Philadelphia*, 1993.

[29] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. of Meetings on Acoustics ICA2013*. Acoustical Society of America, June 2013, vol. 19, p. 035081.

CHAPTER 4

Conclusion

4.1 Summary

Speech enhancement has received widespread attention for decades. Clean speech recorded by microphones is inevitably distorted by interfering noise, degrading speech quality and intelligibility. Depending on the application scenario, various types of noise may be involved. This imposes challenges for speech communication devices to work reliably in noisy environments. Traditional algorithms attempted to track signal power changes in the time-frequency domain, exhibiting limited effectiveness in acoustically challenging conditions. Recently, machine learning algorithms have emerged as the prevailing technique with noticeable performance improvements, especially these DNN-based predictive approaches. To achieve superior performance, these algorithms often require training DNNs on large amounts of labeled data covering a wide range of acoustic conditions. However, deep predictive models may struggle to generalize well to unseen acoustic conditions. As a result, the network models may generate erroneous clean speech estimates without providing any indication of uncertainty. Moreover, the lack of transparency into how neural networks generate their estimates makes these models difficult to interpret. This thesis focuses on improving the interpretability and robustness of DNN-based speech enhancement by incorporating statistical models. In this chapter, we summarize the main findings and contributions of the thesis and discuss future research directions.

4.1.1 Uncertainty in Deep Predictive Speech Enhancement

In ***Chapter 2***, we studied these problems from the perspective of uncertainty modeling, aiming to enable network models to provide predictive uncertainty. Rather than solely estimating multiplicative filter masks, we augment supervised masking approaches with an uncertainty estimation task. In [P2], we follow the widely used complex Gaussian prior of speech and noise spectral coefficients and utilize DNNs to estimate the parameters required for the speech posterior. In this case, the posterior distribution of clean speech is also a complex Gaussian parameterized by its mean and variance. The predicted mean serves as a clean speech estimate, and the variance is interpreted as the associated prediction uncertainty. One of the challenges in optimizing DNN-based probabilistic models is

training instability. Indeed, the ground truth of uncertainty is not readily available, making uncertainty estimation an unsupervised task with an unspecified search space. In addition, we intended to use the uncertainty estimates to further improve speech enhancement performance. Finally, we proposed to address these two problems with a unified solution by utilizing a hybrid loss function. Specifically, we incorporated an approximate MAP estimator of spectral magnitudes that explicitly requires the variance estimates of complex speech spectral coefficients, establishing an interesting connection between the complex spectral domain and the magnitude domain in a single deep framework.

The proposed loss function has experimentally shown its effectiveness in stabilizing the training process. Meanwhile, incorporating uncertainty into clean speech estimation via an approximate MAP magnitude estimator leads to improved speech enhancement performance compared to the baselines that produce filter masks only. The performance improvement could be attributed to two factors. First, explicitly considering uncertainty estimates in the magnitude estimator tends to cause less speech attenuation than the Wiener filter, especially for low inputs, as shown in Figure 2 of [P2] in Section 2.1. This potentially improves the robustness compared to the algorithms that estimate only a Wiener filter. The experimental results across a range of SNRs have demonstrated that the speech preservation tendency of the uncertainty-based speech estimator yields a more pronounced improvement in speech quality at high input SNRs. This is because speech quality at high SNRs is mainly affected by speech distortions. Second, the approximated MAP estimator of spectral magnitudes is nonlinear with respect to the noisy input, while a multiplicative mask has a linear relationship with the noisy input. Thus, when the Wiener estimate and associated uncertainty are jointly optimized using the proposed loss function, the DNNs are expected to generate more accurate clean speech estimates, surpassing the simple linear Wiener filter.

We further present a comprehensive analysis of data uncertainty on a time-frequency scale. We hypothesize that uncertainty estimates are supposed to be closely related to estimation errors. In other words, when the predictions deviate greatly from the true data, the network model should output large estimates to indicate that its predictions are uncertain, and vice versa. The quantitative analysis over the time-frequency bins confirmed that the captured data uncertainty accurately reflects estimation errors in predictions. This is of great interest because data uncertainty can serve as an indicator of the degree of confidence we can place in the predictions. Furthermore, accurate uncertainty estimation is essential for the spectral magnitude estimator to operate as intended. This is because the approximate MAP magnitude estimator's value is positively correlated with the uncertainty estimate when other terms in the formula are fixed, see equation (10) in [P2]. When the clean speech estimate deviates largely from the true target, resulting in large errors, the corresponding large uncertainty estimate increases the value of the magnitude estimator, thereby preserving more clean speech. It degenerates into the Wiener filter when uncertainty approaches zero.

Besides data uncertainty, we also explored model uncertainty in deep speech enhancement utilizing Bayesian deep learning techniques. Instead of modeling the DNN's weights deterministically, Bayesian deep learning places a distribution over the network parameters and approximates the posterior distribution of these stochastic parameters. Specifically, our study in [P2] leveraged MC

dropout and Deep ensembles to simultaneously estimate clean speech and capture model uncertainty, primarily due to their scalability to large models and datasets. These methods empirically approximate the predictive distribution by sampling from the network weight posterior distribution during inference. Although increasing the number of sampling times can theoretically lead to more accurate uncertainty estimates, the evaluation results show that the performance of uncertainty estimation may eventually saturate. The experiments also show that compared to MC dropout, Deep ensembles can provide not only more reliable uncertainty estimates but also more accurate clean speech estimates. This may be due to that Deep ensembles trained with different random seeds can exploit the multi-modal nature of the high-dimensional network parameter space, whereas MC dropout may only sample around a single mode.

While generic Bayesian deep learning methods are task-agnostic and can be flexibly applied to a wide range of tasks, they often require an expensive sampling process, which poses challenges to applications with limited computational resources or real-time constraints. Thus, another research topic we explored in the first part is to leverage domain-specific knowledge and enable the network to capture different sources of uncertainty with only a single forward pass during inference, thereby avoiding the computational load of sampling. In [P3], we model the spectral coefficients of speech and noise with CGMMs, which can model the super-Gaussian characteristics of speech spectral coefficients. Unlike other super-Gaussian priors that may yield complex posterior derivations [7, 24], the CGMM assumption reaches an advantageous trade-off between modeling complex distributions and analytical tractability. Similarly, we utilize DNNs to estimate the necessary parameters involved in the speech posterior, which also follows a CGMM. However, two problems may arise when optimizing this probabilistic network model: training instability and mode collapse. In our case, training DNNs to minimize the negative log-posterior has experimentally shown instability due to the dependence of the gradient on the variance. To address this issue, we employed and adapted a gradient modification scheme. Moreover, it is indeed challenging to estimate the parameters of the multiple Gaussian components in the CGMM based on a single input observation. The DNN may take shortcuts to output an identical set of estimates, leading to the mode collapse problem. To address this issue, we leverage a winner-take-all pre-training scheme to promote diverse predictions.

We can interpret network models that output the CGMM parameters as generating multiple hypotheses. This can be particularly advantageous when dealing with difficult inputs, where inferring a single accurate estimate might be difficult. In contrast, a model that can make multiple guesses would increase the likelihood that an accurate estimate lies in the output space [104], thus potentially improving the model's robustness. The experimental results on different datasets and various SNRs in [P3] have also confirmed the benefits of multi-hypothesis modeling over a comparable baseline inferring a single best prediction. Moreover, these multiple Gaussian estimates in a CGMM can simultaneously offer data uncertainty and model uncertainty estimates, thus circumventing the need for an expensive sampling process. Qualitative and quantitative analyses have been performed to show the correlation between estimation errors and predictive uncertainties. The close distance of the sparsification plots to the ground truth indicates the high reliability of the uncertainty estimates. As

a result, incorporating domain-specific statistical modeling into deep predictive approaches enables effective and efficient uncertainty modeling. Additionally, statistical modeling acting as a form of regularization can further boost speech enhancement performance, thereby improving the overall interpretability and robustness of the framework.

One straightforward way to address generalization issues is to retrain the DNN using newly collected data from the target domain. However, the process of collecting paired data is often cumbersome and may require costly post-processing. In contrast, gathering unlabeled noisy mixtures is more practical, and over time, the quantity of such noisy data can reach a considerable size. Therefore, in [P4], we investigated methods to mitigate performance degradation caused by domain mismatch using only noisy mixtures, achieving unsupervised domain adaptation. For this, we followed the idea of remixing-based domain adaptation, where a student model is fine-tuned to approximate the pseudo-targets generated by the teacher model. However, when the teacher model's estimates are erroneous, the student model is forced to align with fundamentally incorrect pseudo-targets. To improve the data quality, we first enable the teacher model to estimate the uncertainty associated with its clean speech predictions. This is followed by filtering out low-quality pseudo-targets generated by the teacher model in the target domain, based on the uncertainty estimates, such that the student model learns from only high-quality speech estimates. Eventually, the uncertainty-based remixing allows the student model to effectively capture noise characteristics of the target domain, leading to better noise reduction capability at the cost of some speech distortion. Additionally, adjusting the uncertainty threshold can achieve a controllable trade-off between noise reduction and speech distortion.

Achieving high robustness across different acoustic conditions based on a single modality has been challenging. This problem can be alleviated by leveraging features from other modalities. For example, machines understand speech via recognition systems, whose performance can be negatively affected by severe noise interference. However, complementary features from other modalities, such as lip movements extracted from visual signals, can overcome the limitations since it is independent of acoustic corruptions. This gives rise to the task of audio-visual phoneme recognition. An important question involved in multi-modal methods is how to effectively fuse information from multiple modalities. Moreover, similar to noisy audio inputs, video data is not guaranteed to be consistently informative possibly due to issues such as object occlusion and illumination conditions. Unreliable visual inputs may provide misleading information, resulting in degraded performance that may be even worse than methods based solely on audio modality. Thus, the research question we investigated in [P5] is how to improve the model's robustness to simultaneously corrupted video and audio. For this, we proposed to guide the fusion of audio and visual information by incorporating modality-wise uncertainty, based on which we determine to which extent the final decision can rely on each modality. In contrast to work in [126], which takes certain visual corruptions into the training process and consequently may only perform well on in-distribution corrupted inputs, our uncertainty-based fusion scheme is corruption-agnostic, which is expected to generalize well to a variety of possible noisy video inputs. Our experimental results have demonstrated that the proposed hybrid fusion scheme is robust

under noisy audio-visual conditions, while at the same time, still making use of the complementary advantages of multi-modal methods when the video is sufficiently clean.

In Chapter 2, we have mainly discussed uncertainty modeling in the context of deep speech enhancement. This includes capturing uncertainty based on different statistical assumptions and exploring its use in unsupervised domain adaptation and in multi-modal information fusion. Uncertainty modeling makes the DNN-based algorithms more interpretable and can be achieved at negligible cost. We believe uncertainty modeling is an important feature of deep learning algorithms.

Research Questions

RQ1 *How can uncertainty be effectively modeled in deep predictive speech enhancement, and to what extent can uncertainty estimates reliably predict deviations from ground-truth speech? How does uncertainty estimation affect speech enhancement performance?*

In [P2], we follow the widely used complex Gaussian priors of speech and noise spectral coefficients and estimate the full clean speech posterior using DNNs, where the speech posterior variance serves as a data uncertainty measure. We further embedded this into Bayesian deep learning to additionally estimate model uncertainty. We present a comprehensive analysis of uncertainty estimates, both qualitatively and quantitatively, demonstrating that the uncertainty-augmented speech enhancement model can inform us of incorrect clean speech estimates through increased associated uncertainty estimates. Furthermore, we propose a hybrid loss function for training DNNs that achieves two objectives. First, it stabilizes the training of the DNN-based probabilistic model. Second, it integrates uncertainty estimates into a statistically principled uncertainty-aware speech estimator, yielding superior speech enhancement performance than the baseline that uses the same architecture to output single-point estimates. Combining the proposed training loss with Bayesian deep learning frameworks can further improve the performance.

RQ2 *How can one leverage statistical domain knowledge to develop more efficient methods for uncertainty estimation in deep predictive speech enhancement?*

In [P3], we model the spectral coefficients of speech and noise with CGMMs, which can potentially approximate any probability density with arbitrary accuracy. This offers a good fit to model super-Gaussian characteristics of speech spectral coefficients and ensures a relatively straightforward posterior derivation. We predict the clean speech posterior resulting from the CGMM priors using a DNN. We demonstrate that the proposed approach can efficiently predict both data uncertainty and model uncertainty with only a single forward pass of the DNN. Furthermore, we show that the proposed uncertainty-augmented approach that combines powerful statistical models and deep learning also delivers a superior speech enhancement performance compared to the method that estimates single-point clean speech estimates.

Research Questions

RQ3 *Can uncertainty estimates be further leveraged to improve the robustness and generalization ability of speech enhancement systems?*

We introduce in [P4] a uncertainty-based remixing strategy to mitigate performance degradation caused by domain mismatch of data. Our experimental results demonstrate that the proposed adaptation method provides improved generalization performance compared to the baseline trained only on data from the source domain. Additionally, adjusting the uncertainty threshold allows for an interesting controllable trade-off between noise reduction and speech distortion. We propose in [P5] a uncertainty-driven multi-modal information fusion strategy for audio-visual phoneme recognition. We observe that the proposed uncertainty-driven fusion scheme performs better on simultaneously corrupted audio and visual inputs than the baseline using the intermediate feature fusion scheme, demonstrating improved robustness to unseen modality corruption. Meanwhile, it maintains the benefits offered by multi-modal methods when the input is sufficiently informative.

4.1.2 Noise-Aware Generative Speech Enhancement Based on Variational Autoencoder and Non-Negative Matrix Factorization

In ***Chapter 3***, we explored the interpretability and generalization issues by focusing on deep generative methods, specifically the VAE-NMF framework [21, 20]. This framework offers statistical interpretability and elegantly embeds DNNs into a statistical framework for speech enhancement. In contrast to predictive approaches learning deterministic mapping relationships between noisy mixtures and clean speech, such semi-supervised generative methods learn a prior distribution over clean speech and reuse this knowledge to extract clean speech from noisy mixtures. Deep generative approaches can potentially generalize to various acoustic conditions by focusing on learning the underlying distribution of data. However, training the network model on isolated clean speech data does not always ensure high robustness in challenging acoustic environments. For instance, these methods may struggle to perform robustly when both ego-noise and environmental noise are present. This may occur when the target speaker talks to a moving robot in a noisy environment. In this chapter, we improve the robustness of the deep generative methods further, while retaining the statistical interpretability. Specifically, we incorporated the noise information into the speech [P6] and noise [P7] [P8] models, respectively.

In [P6], we sought to improve the speech model by developing a noise-aware encoder for the VAE. Compared with the vanilla VAE aiming to discover the latent variables of clean speech by training the network model on clean speech only, we proposed to infer the latent variables of noisy speech and refine them by closing the difference to that of the corresponding clean speech [P6]. This is achieved by a two-step training strategy: The vanilla VAE is trained on clean speech only first and

then used to guide the training of the noise-aware encoder in the latent space. The refinement step is performed only on the encoder part while keeping the decoder obtained on clean speech unchanged. This noise-aware speech model is then combined with a NMF-based noise model whose unknown parameters are obtained by performing maximum likelihood estimation. The experimental results have shown that the noise-aware training does not deteriorate the method's generalization capabilities but yields better speech enhancement performance without introducing any additional computational efforts.

In [P7] and [P8], we proposed to augment the noise model by incorporating the prior knowledge from human-robot interaction. Since ego-noise exhibits structured patterns in the time-frequency domain, such as harmonic structures, we incorporated these temporal-spectral characteristics by pre-training a separate NMF model on isolated ego-noise recordings. This is then combined with an adaptive NMF model to account for unseen environmental noise. Eventually, the unknown parameters are jointly optimized to find a maximum likelihood solution during inference. We evaluate this partially adaptive scheme on various acoustic conditions, including different SNRs and interaction scenarios. The experimental results over different instrumental metrics have shown that the performance improvement is most pronounced in the most challenging case, where speech is distorted by ego-noise and environmental noise simultaneously. In comparison, the fully adaptive scheme has difficulty in capturing non-stationary ego-noise properly, while the fully learned and fixed scheme struggles to model unseen environmental noise. Furthermore, the partially adaptive concept can be extended to the multichannel cases, where the pre-training strategy leverages not only the temporal-spectral characteristics of ego-noise but also its spatial information. Unlike environmental noise, whose characteristics are often unknown in advance and closely related to the surrounding environment, ego-noise is mainly generated by the motors distributed over the robot's body. Due to the fact that the joints can only move with limited degrees of freedom, ego-noise also exhibits limited spatial diversity, which can be effectively captured by the multichannel partial adaptive scheme. Therefore, we proposed in [P8] to extend the multichannel VAE-NMF framework to incorporate both spatial and spectral knowledge of ego-noise. This is achieved by fixing both the learned dictionary matrix and spatial covariance matrix during inference. The proposed partially adaptive multi-channel framework shows similar performance improvement trends as the single-channel extension: In the presence of joint ego-noise and environmental noise distortions, it yields superior performance compared to the baselines, which include a fully adaptive multichannel scheme incorporating no prior noise information and a fully learned and fixed multichannel scheme without adaptation capabilities. Overall, the proposed noise-aware partially adaptive scheme can effectively suppress ego-noise, while at the same time adapting to unseen environmental noise.

The VAE-NMF framework excels in generalizing to unseen noisy mixtures, showcasing the potential of deep generative methods in speech enhancement. The experiments have shown that incorporating prior noise information alongside clean speech during training is crucial to achieving high robustness under severe noise interference.

Research Questions

RQ4 *How can VAE-NMF-based generative approaches leverage prior noise knowledge to improve speech modeling capabilities for better generalization in unseen acoustic environments?*

We observe that the VAE-based speech model that learns a prior distribution of clean speech faces the challenge of performing robustly in the presence of noise. Thus, we introduce a noise-aware encoder for the VAE model. The method refines the latent variables inferred from noisy speech by narrowing down the difference with the latent variables of the corresponding clean speech. The resulting method preserves the original statistical framework's interpretability. The experimental evaluation demonstrates that the proposed noise-aware speech model exhibits improved generalization ability to unseen noise compared to the speech model trained only on clean speech using the same architecture.

RQ5 *Can one derive a flexible and effective noise adaptation scheme that can reuse learned noise representation while adapting to unseen noise characteristics? Furthermore, can such an adaptation scheme be extended to multichannel applications?*

We observe that the fully adaptive noise learning scheme can potentially generalize to various acoustic conditions but has difficulty performing robustly in challenging noisy environments, e.g., involving both ego-noise and environmental noise. In contrast, the learned and fixed scheme may perform well on seen acoustic conditions but fails to capture unseen noise characteristics. Therefore, in [P7], we propose a partially adaptive noise learning scheme that leverages prior representations learned from the collected ego-noise while adapting to unseen environmental noise characteristics, with unknown parameters derived jointly under the maximum likelihood criterion. This partially adaptive scheme is further extended to multichannel application scenarios in [P8]. We demonstrate that the resulting noise adaptation approach surpasses baselines that either learn noise features on the fly or rely solely on learned prior noise representation.

4.2 Discussion and Future Work

This thesis focuses on DNN-based speech enhancement. While it provides excellent performance across a range of application scenarios, there are also some costs. For instance, high performance often comes with a significant increase in the number of parameters, which imposes great challenges on resource-limited devices. Although research on designing lighter and more efficient network architectures has received increasing attention, a general trend can be observed that large networks can more easily outperform small networks. Therefore, statistical regularization might be particularly beneficial for lightweight networks, which are considered more challenging to gain high performance.

Rather than learning a lightweight model from scratch, another promising research direction is to use knowledge distillation, which utilizes a teacher-student framework. Unlike the one used in our domain adaptation problem in [P4], the teacher model used in knowledge distillation is generally larger and more powerful than the student. In such a way, the teacher network can transfer some abstract knowledge that can not be easily learned by smaller networks to the student model. Knowledge distillation has become increasingly popular in deep learning and computer vision since the work by Hinton et al. [179]. Moreover, this learning paradigm is also supported by theoretical perspectives grounded in probability theory [179]. For example, a teacher classifier may output similar probability values for pictures of a baby tiger and a baby cat that looks like a tiger. From the perspective of entropy, this set of probabilities is more informative and has higher entropy compared to one-hot encoding that tells you exactly if it belongs to a cat or tiger. As a result, training a small student network with this pseudo target may outperform the same model architecture learning from scratch. However, most of its applications are tailored for classification settings, and it is not straightforward to directly transfer the theory applicable to classification tasks to regression settings. Nevertheless, existing work has attempted to apply the knowledge distillation to achieve tiny speech enhancement [180, 181]. In such cases, the powerful teacher model can provide pseudo-clean targets to train a student model. However, it raises the same problem that the teacher model may provide misleading guidance to the student model when its predictions are inaccurate as we discussed in unsupervised domain adaptation [P4]. To solve this, it can naturally be combined with uncertainty modeling, where the teacher model selectively transfers reliable knowledge to the student model based on its uncertainty estimates. Another interesting research direction could be to combine teacher-student learning with uncertainty-based curriculum learning. More specifically, we can employ an uncertainty-based teacher model to progressively guide the student model's training. This can be achieved by ranking the difficulty of unlabeled data based on uncertainty assigned by the teacher model to the corresponding clean speech estimates, such that a student model can be trained on samples ranging from easy to difficult.

While in this thesis we evaluated the proposed uncertainty-based multi-modality fusion in the task of audio-visual speech recognition, it can also be extended to various other tasks, such as audio-visual emotion recognition and event detection. Furthermore, this is not necessarily limited to audio and video modalities but can be extended to others such as text or biomedical signals. This is because the proposed uncertainty-based fusion is designed to be independent of the specific type of modality, but focuses on leveraging modality-wise confidences to perform modality fusion effectively. More interestingly, extending the uncertainty-based fusion to encompass more than two modalities may have a broader impact on practical applications. For this, future work may include replacing the current engineered fusion scheme with a DNN-based learnable uncertainty-based fusion strategy.

The latter part of the thesis focuses on deep generative methods, specifically, the VAE-NMF framework, which combines DNNs and statistical modeling in a principled way. However, the optimization strategy at inference requires a costly MC sampling process, which may not meet real-time constraints. Further research may focus on improving its inference speed. Recent work has pushed this direction forward, such as leveraging variational inference [62] and Lagrange dynamics [182]. Developing a

more effective and faster inference scheme remains a promising area for further exploration. Another computationally expensive factor during inference is the iterative updating required to obtain the NMF parameters. Moreover, NMF reconstructs the input by a linear combination of templates, and the inherent linearity in the reconstruction process may potentially restrict the overall performance. This raises another question: Can we replace the NMF model with other techniques to avoid this computational burden and modeling limitation? Existing work has attempted to find a neural network-based alternative by injecting the non-negative concept into an auto-encoder architecture [183]. Further research can be conducted to combine it with the VAE-based speech model. Additionally, maintaining its adaptation ability to generalize well to unseen acoustic conditions would be an interesting research topic to explore next.

Recent work has shown that the research interest in deep generative models has expanded to include diffusion models. Diffusion models can be trained to learn the probability distribution of complex data, such as text, image, and speech. Unlike VAE, diffusion models do not necessarily need to represent input data through a low-dimensional latent space, which may result in the loss of fine details. Additionally, the diffusion-based speech enhancement methods can be designed to fully leverage the non-linear modeling capabilities of DNNs, without the need to consider a NMF-based noise model [86]. An interesting future research direction could focus on adapting the technique to make it specific to the speech domain, e.g., by incorporating statistical knowledge of speech and noise signals.

References

[1] A. Schmidt, H. W. Löllmann, and W. Kellermann, "Acoustic self-awareness of autonomous systems in a world of sounds," *Proceedings of the IEEE*, vol. 108, no. 7, pp. 1127–1149, 2020.

[2] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, "Mechatronic design of NAO humanoid," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, Kobe, Japan, 2009, pp. 769–774.

[3] A. Robotics, "NAOqi documentation center," http://doc.aldebaran.com/.

[4] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[5] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement*. Springer, 2013.

[6] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Proc.*, vol. 9, no. 5, pp. 504–512, 2001.

[8] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Infor. Proc. Systems (NeurIPS)*, Denver, Colorado, USA, 2000, p. 535–541.

[9] A. Deleforge and W. Kellermann, "Phase-optimized K-SVD for signal extraction from under-determined multichannel sparse mixtures," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Brisbane, Australia, 2015, pp. 355–359.

[10] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 21, no. 5, pp. 971–982, 2013.

[11] K. P. Murphy, *Probabilistic machine learning: Advanced topics*. MIT Press, 2023. [Online]. Available: http://probml.github.io/book2

[12] R. Rehr and T. Gerkmann, "SNR-based features and diverse training data for robust DNN-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 29, pp. 1937–1949, 2021.

[13] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in *ISCA Interspeech*, Stockholm, Sweden, 2017, pp. 1993–1997.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Munich, Germany, 2015, pp. 234–241.

[15] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 28, pp. 1778–1787, 2020.

[16] Y. Fu, Y. Liu, J. Li, D. Luo, S. Lv, Y. Jv, and L. Xie, "Uformer: A UNet based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Singapore, Singapore, 2022, pp. 7417–7421.

[17] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *International Conf. on Telecommunications and Signal Processing (TSP)*, Virtual, 2021, pp. 72–76.

[18] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *ISCA Interspeech*, Stockholm, Sweden, 2017, pp. 3642–3646.

[19] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chén, P. Koch, M. De Vos, and A. Mertins, "Improving GANs for speech enhancement," *IEEE Signal Proc. Letters*, vol. 27, pp. 1700–1704, 2020.

[20] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *IEEE Int. Workshop on Machine Learning for Signal Proc. (MLSP)*, Aalborg, Denmark, 2018, pp. 1–6.

[21] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Calgary, Alberta, Canada, 2018, pp. 716–720.

[22] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 30, pp. 2993–3007, 2022.

[23] T. Virtanen, E. Vincent, and S. Gannot, "Time-frequency processing: Spectral properties," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. Wiley, 2018, pp. 15–29.

[24] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 15, no. 6, pp. 1741–1752, 2007.

[25] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. Wiley, 2018, pp. 65–85.

[26] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*. IGI global, 2011, pp. 162–185.

[27] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[28] J. S. Erkelens, R. C. Hendriks, and R. Heusdens, "On the estimation of complex speech DFT coefficients without assuming independent real and imaginary parts," *IEEE Signal Proc. Letters*, vol. 15, pp. 213–216, 2008.

[29] R. Martin, "Speech enhancement based on minimum mean-square error estimation and super-Gaussian priors," *IEEE Trans. on Speech and Audio Proc.*, vol. 13, no. 5, pp. 845–856, 2005.

[30] ——, "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Orlando, Florida, USA, 2002, pp. 253–256.

[31] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. on Adv. in Signal Proc.*, vol. 2005, pp. 1–17, 2005.

[32] R. F. Astudillo, "An extension of STFT uncertainty propagation for GMM-based super-Gaussian a priori models," *IEEE Signal Proc. Letters*, vol. 20, no. 12, pp. 1163–1166, 2013.

[33] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006.

[34] S. Markovich-Golan, W. Kellermann, and S. Gannot, "Spatial filtering," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. Wiley, 2018, pp. 191–212.

[35] N. Q. Duong, E. Vincent, and R. Gribonval, "Spatial covariance models for under-determined reverberant audio source separation," in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, New Paltz, NY, USA, 2009, pp. 129–132.

[36] S. Markovich-Golan, W. Kellermann, and S. Gannot, "Multichannel parameters estimation," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. Wiley, 2018, pp. 219–231.

[37] M. I. Mandel, S. Araki, and T. Nakatani, "Multichannel clustering and classification approaches," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. Wiley, 2018, pp. 235–258.

[38] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Brighton, UK, 2019, pp. 101–105.

[39] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Proc. Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[40] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Dallas, Texas, USA, 2010, pp. 4266–4269.

[41] C. Févotte, N. Bertin, and J.-L. Durrieu, "Non-negative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[42] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[43] E. Vincent, S. Gannot, and T. Virtanen, "Introduction," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. Wiley, 2018, pp. 3–12.

[44] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Infor. Proc. Systems (NeurIPS)*, Vancouver, Online, Canada, 2020, pp. 596–608.

[45] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 28, pp. 1404–1415, 2020.

[46] A. Nicolson and K. K. Paliwal, "Masked multi-head self-attention for causal speech enhancement," *Speech Communication*, vol. 125, pp. 80–96, 2020.

[47] Q. Zhang, Q. Song, A. Nicolson, T. Lan, and H. Li, "Temporal convolutional network with frequency dimension adaptive attention for speech enhancement," in *ISCA Interspeech*, Brno, Czech Republic, 2021, pp. 166–170.

[48] A. Nicolson and K. K. Paliwal, "On training targets for deep learning approaches to clean speech magnitude spectrum estimation," *The Journal of the Acoustical Society of America*, vol. 149, no. 5, pp. 3273–3293, 2021.

[49] M. Tammen, D. Fischer, B. T. Meyer, and S. Doclo, "DNN-based speech presence probability estimation for multi-frame single-microphone speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 191–195.

[50] Y. A. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 20, no. 4, pp. 1256–1269, 2011.

[51] M. Tammen and S. Doclo, "Deep multi-frame MVDR filtering for single-microphone speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 8443–8447.

[52] D. Fischer and S. Doclo, "Sensitivity analysis of the multi-frame MVDR filter for single-microphone speech enhancement," in *Proc. Euro. Signal Proc. Conf. (EUSIPCO)*, Kos Island, Greece, 2017, pp. 603–607.

[53] A. Schasse and R. Martin, "Estimation of subband speech correlations for noise reduction via MVDR processing," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 22, no. 9, pp. 1355–1365, 2014.

[54] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, D. S. Williamson, and D. Yu, "Multi-channel multi-frame ADL-MVDR for target speech separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 29, pp. 3526–3540, 2021.

[55] Y. G. Jin, J. W. Shin, and N. S. Kim, "Spectro-temporal filtering for multichannel speech enhancement in short-time Fourier transform domain," *IEEE Signal Proc. Letters*, vol. 21, no. 3, pp. 352–355, 2014.

[56] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[57] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[58] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Shanghai, China, 2016, pp. 31–35.

[59] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Brisbane, Australia, 2015, pp. 708–712.

[60] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 24, no. 3, pp. 483–492, 2015.

[61] T. Peer and T. Gerkmann, "Phase-aware deep speech enhancement: It's all about the frame length," *J. Acoust. Soc. Am.*, vol. 2, no. 10, 2022.

[62] M. Pariente, "Implicit and explicit phase modeling in deep learning-based source separation," Ph.D. dissertation, Université de Lorraine, 2021.

[63] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Proc. Letters*, vol. 32, no. 2, pp. 55–66, 2015.

[64] J.-M. Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement," in *International Workshop on Multimedia Signal Processing (MMSP)*, Vancouver, BC, Canada, 2018, pp. 1–5.

[65] D. Ditter and T. Gerkmann, "A multi-phase gammatone filterbank for speech separation via TasNet," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 36–40.

[66] R. D. Nathoo, M. Kegler, and M. Stamenovic, "Two-step knowledge distillation for tiny speech enhancement," *arXiv preprint arXiv:2309.08144*, 2023.

[67] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Separated noise suppression and speech restoration: LSTM-based speech enhancement in two stages," in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, New Paltz, NY, USA, 2019, pp. 239–243.

[68] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A dissecting approach," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 6359–6363.

[69] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[70] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Calgary, Alberta, Canada, 2019, pp. 626–630.

[71] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *IEEE Workshop Autom. Speech Recog. and Underst. (ASRU)*, Scottsdale, AZ, USA, 2015, pp. 444–451.

[72] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[73] K. Alex, B. Vijay, and C. Roberto, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *Proc. of the British Machine Vision Conf. (BMVC)*, London, UK, 2017, pp. 57.1–57.12.

[74] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Advances in Neural Infor. Proc. Systems (NeurIPS)*, Long Beach, CA, USA, 2017, p. 5580–5590.

[75] F. K. Gustafsson, M. Danelljan, and T. B. Schon, "Evaluating scalable Bayesian deep learning methods for robust computer vision," in *IEEE/CVF Conf. on Computer Vision and Pattern Recog. (CVPR)*, Seattle, Washington, USA, 2020, pp. 318–319.

[76] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Int. Society for Music Info. Retrieval Conf. (ISMIR)*, Suzhou, China, 2017.

[77] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *ISCA Interspeech*, Hyderabad, India, 2018, pp. 3229–3233.

[78] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 656–660.

[79] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 29, pp. 2001–2014, 2021.

[80] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-path convolution recurrent network for single channel speech enhancement," in *ISCA Interspeech*, Brno, Czech Republic, 2021, pp. 2811–2815.

[81] J. Wang, Z. Lin, T. Wang, M. Ge, L. Wang, and J. Dang, "Mamba-SEUNet: Mamba UNet for monaural speech enhancement," in *ICASSP*, Hyderabad, India, 2025.

[82] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "SNR-based progressive learning of deep neural network for speech enhancement," in *ISCA Interspeech*, San Francisco, CA, USA, 2016, pp. 3713–3717.

[83] X. Hao, X. Su, S. Wen, Z. Wang, Y. Pan, F. Bao, and W. Chen, "Masking and inpainting: A two-stage speech enhancement approach for low SNR and non-stationary noise," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 6959–6963.

[84] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Int. Conf. Machine Learning (ICML)*, Long Beach, California, USA, 2019, pp. 2031–2041.

[85] Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually guided speech enhancement using deep neural networks," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Calgary, Alberta, Canada, 2018, pp. 5074–5078.

[86] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, 2023.

[87] D. de Oliveira, S. Welker, J. Richter, and T. Gerkmann, "The PESQetarian: On the relevance of Goodhart's law for speech enhancement," in *Interspeech*, Kos Island, Greece, 2024, pp. 3854–3858.

[88] 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.

[89] A. Pandey and D. Wang, "Self-attending RNN for speech enhancement to improve cross-corpus generalization," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 30, pp. 1374–1385, 2022.

[90] M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Filterbank design for end-to-end speech separation," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 6364–6368.

[91] D. de Oliveira, T. Peer, and T. Gerkmann, "Efficient transformer-based speech enhancement using long frames and STFT magnitudes," in *ISCA Interspeech*, Incheon, Korea, 2022, pp. 2948–2952.

[92] S. Lv, Y. Hu, S. Zhang, and L. Xie, "DCCRN+: Channel-wise subband DCCRN with SNR estimation for speech enhancement," *arXiv preprint arXiv:2106.08672*, 2021.

[93] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *ISCA Interspeech*, Shanghai, China, 2020, pp. 2472–2476.

[94] H. Wu, K. Tan, B. Xu, A. Kumar, and D. Wong, "Rethinking complex-valued deep neural networks for monaural speech enhancement," in *ISCA Interspeech*, Dublin, Ireland, 2023, pp. 3889–3893.

[95] F. E. R. Ilg, "Estimating optical flow with convolutional neural networks," Ph.D. dissertation, Albert-Ludwigs-Universität Freiburg, 2020.

[96] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning," in *Int. Conf. Machine Learning (ICML)*, Stockholm, Sweden, 2018, pp. 1184–1193.

[97] T. Pearce, A. Brintrup, M. Zaki, and A. Neely, "High-quality prediction intervals for deep learning: A distribution-free, ensembled approach," in *Int. Conf. Machine Learning (ICML)*, Stockholm, Sweden, 2018, pp. 4075–4084.

[98] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Infor. Proc. Systems (NeurIPS)*, Long Beach, CA, USA, 2017, p. 6405–6416.

[99] S. Depeweg, "Modeling epistemic and aleatoric uncertainty with Bayesian neural networks and latent variables," Ph.D. dissertation, Technische Universität München, 2019.

[100] G. Xu, Z. Liu, and C. C. Loy, "Computation-efficient knowledge distillation via uncertainty-aware mixup," *Pattern Recognition*, vol. 138, p. 109338, 2023.

[101] A. Malinin, "Uncertainty estimation in deep learning with application to spoken language assessment," Ph.D. dissertation, University of Cambridge, 2019.

[102] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?" *Structural safety*, vol. 31, no. 2, pp. 105–112, 2009.

[103] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine Learning*, vol. 110, pp. 457–506, 2021.

[104] E. Ilg, O. Cicek, S. Galesso, A. Klein, O. Makansi, F. Hutter, and T. Brox, "Uncertainty estimates and multi-hypotheses networks for optical flow," in *European Conf. on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 652–667.

[105] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP J. on Adv. in Signal Proc.*, vol. 2003, no. 10, pp. 1–9, 2003.

[106] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *Int. Conf. Machine Learning (ICML)*, Bellevue, Washington, USA, 2011, pp. 681–688.

[107] T. Chen, E. Fox, and C. Guestrin, "Stochastic gradient Hamiltonian Monte Carlo," in *Int. Conf. Machine Learning (ICML)*, Beijing, China, 2014, pp. 1683–1691.

[108] Y.-A. Ma, T. Chen, and E. Fox, "A complete recipe for stochastic gradient MCMC," in *Advances in Neural Infor. Proc. Systems (NeurIPS)*, Montreal, Quebec, Canada, 2015, p. 2917–2925.

[109] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Int. Conf. Learning Repr. (ICLR)*, Y. Bengio and Y. LeCun, Eds., Banff, Canada, 2014.

[110] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Int. Conf. Machine Learning (ICML)*, Lille, France, 2015, pp. 1613–1622.

[111] M.-N. Tran, T.-N. Nguyen, and V.-H. Dao, "A practical tutorial on variational Bayes," *arXiv preprint arXiv:2103.01327*, 2021.

[112] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[113] N. R. 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 Trans. on Affective Computing*, pp. 1–14, 2023.

[114] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Int. Conf. Machine Learning (ICML)*, New York, NY, USA, 2016, pp. 1050–1059.

[115] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[116] S. Braun and S.-C. Liu, "Parameter uncertainty for end-to-end speech recognition," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Brighton, UK, 2019, pp. 5636–5640.

[117] S. Khurana, N. Moritz, T. Hori, and J. L. Roux, "Unsupervised domain adaptation for speech recognition via uncertainty driven self-training," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 6553–6557.

[118] A. Vyas, P. Dighe, S. Tong, and H. Bourlard, "Analyzing uncertainties in speech recognition using dropout," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Brighton, UK, 2019, pp. 6730–6734.

[119] K. Sridhar and C. Busso, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 8384–8388.

[120] A. G. Wilson and P. Izmailov, "Bayesian deep learning and a probabilistic perspective of generalization," in *Advances in Neural Infor. Proc. Systems (NeurIPS)*, vol. 33, Vancouver, Online, Canada, 2020, pp. 4697–4708.

[121] S. Fort, H. Hu, and B. Lakshminarayanan, "Deep ensembles: A loss landscape perspective," in *"Bayesian Deep Learning" Neural Information Proc. Systems workshop*, Vancouver, Canada, 2019.

[122] S. Lee, S. Purushwalkam, M. Cogswell, D. Crandall, and D. Batra, "Why M heads are better than one: Training a diverse ensemble of deep networks," *arXiv preprint arXiv:1511.06314*, 2015.

[123] J. Nixon, B. Lakshminarayanan, and D. Tran, "Why are bootstrapped deep ensembles not better?" in *"I Can't Believe It's Not Better!" Neural Information Proc. Systems workshop*, Vancouver, Online, Canada, 2020.

[124] N. Raj Prabhu, G. Carbajal, N. Lehmann-Willenbrock, and G. Timo, "End-to-end label uncertainty modeling for speech-based arousal recognition using Bayesian neural networks," in *ISCA Interspeech*, Incheon, Korea, 2022, pp. 151–155.

[125] L. Smith and Y. Gal, "Understanding measures of uncertainty for adversarial example detection," in *Uncertainty in Artificial Intelligence (UAI)*, Monterey, CA, USA, 2018, pp. 560–569.

[126] J. Hong, M. Kim, J. Choi, and Y. M. Ro, "Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring," in *IEEE/CVF Conf. on Computer Vision and Pattern Recog. (CVPR)*, Vancouver, Canada, 2023, pp. 18 783–18 794.

[127] W. Yu, S. Zeiler, and D. Kolossa, "Reliability-based large-vocabulary audio-visual speech recognition," *Sensors*, vol. 22, no. 15, p. 5501, 2022.

[128] ——, "Fusing information streams in end-to-end audio-visual speech recognition," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 3430–3434.

[129] Y. Bando, K. Sekiguchi, and K. Yoshii, "Adaptive neural speech enhancement with a denoising variational autoencoder," in *ISCA Interspeech*, Shanghai, China, 2020, pp. 2437–2441.

[130] J. Le Roux, F. J. Weninger, and J. R. Hershey, "Sparse NMF–half-baked or well done?" in *Mitsubishi Electric Research Labs (MERL), Tech. Rep. TR2015-023*, vol. 11, Cambridge, MA, USA, 2015, pp. 13–15.

[131] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Advances in Neural Infor. Proc. Systems (NeurIPS)*, vol. 22, Vancouver, B.C., Canada., 2009.

[132] P. López-Serrano, C. Dittmar, Y. Özer, and M. Müller, "NMF toolbox: Music processing applications of nonnegative matrix factorization," in *International Conf. on Digital Audio Effects (DAFx-19)*, Birmingham, UK, 2019, pp. 2–6.

[133] A. N. Langville, C. D. Meyer, R. Albright, J. Cox, and D. Duling, "Algorithms, initializations, and convergence for the nonnegative matrix factorization," *arXiv preprint arXiv:1407.7299*, 2014.

[134] A. Cichocki, R. Zdunek, and S.-i. Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *International conf. on independent component analysis and signal separation*, Charleston, SC, USA, 2006, pp. 32–39.

[135] S. Sra and I. Dhillon, "Generalized nonnegative matrix approximations with Bregman divergences," in *Advances in Neural Infor. Proc. Systems (NeurIPS)*, vol. 18, Vancouver, Canada, 2005.

[136] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational intelligence and neuroscience*, vol. 2009, no. 1, p. 785152, 2009.

[137] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 18, no. 3, pp. 550–563, 2009.

[138] B. J. King and L. Atlas, "Single-channel source separation using complex matrix factorization," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 19, no. 8, pp. 2591–2597, 2011.

[139] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised speech enhancement based on multichannel NMF-informed beamforming for noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 27, no. 5, pp. 960–971, 2019.

[140] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *ISCA Interspeech*, Makuhari, Japan, 2010, pp. 717–720.

[141] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 21, no. 10, pp. 2140–2151, 2013.

[142] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback–Leibler divergence," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 23, no. 7, pp. 1233–1242, 2015.

[143] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *International Conference on Latent Variable Analysis and Signal Separation*, Tel Aviv, Israel, 2012, pp. 322–329.

[144] P. D. O'Grady and B. A. Pearlmutter, "Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint," *Neurocomputing*, vol. 72, no. 1-3, pp. 88–101, 2008.

[145] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Prague, Czech Republic, 2011, pp. 17–20.

[146] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *IEEE Int. Workshop on Machine Learning for Signal Proc. (MLSP)*, Thessaloniki, Greece, 2007, pp. 431–436.

[147] A. Schmidt, A. Deleforge, and W. Kellermann, "Ego-noise reduction using a motor data-guided multichannel dictionary," in *International Conf. on Intelligent Robots and Systems (IROS)*, Daejeon, Korea, 2016, pp. 1281–1286.

[148] T. Haubner, A. Schmidt, and W. Kellermann, "Multichannel nonnegative matrix factorization for ego-noise suppression," in *Speech Communication; 13th ITG-Symposium*, Oldenburg, Germany, 2018, pp. 1–5.

[149] A. Schmidt, A. Brendel, T. Haubner, and W. Kellermann, "Motor data-regularized nonnegative matrix factorization for ego-noise suppression," *EURASIP Journal on Audio, Speech, and Music Proc.*, vol. 2020, no. 1, p. 11, 2020.

[150] A. Schmidt and W. Kellermann, "Multichannel nonnegative matrix factorization with motor data-regularized activations for robust ego-noise suppression," in *IEEE International Conf. on Autonomous Systems (ICAS)*, Montréal, Québec, Canada, 2021, pp. 1–5.

[151] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Int. Conf. on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 414–421.

[152] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising." in *ISCA Interspeech*, Brisbane, Australia, 2008, pp. 411–414.

[153] D. P. Kingma, M. Welling *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[154] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, pp. 229–256, 1992.

[155] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Int. Conf. Machine Learning (ICML)*, Beijing, China, 2014, pp. 1278–1286.

[156] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in Neural Infor. Proc. Systems (NeurIPS)*, vol. 27, Montreal, Quebec, Canada, 2014.

[157] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Infor. Proc. Systems (NeurIPS)*, vol. 28, Montreal, Canada, 2015.

[158] E. Aksan and O. Hilliges, "STCN: Stochastic temporal convolutional networks," in *Int. Conf. Learning Repr. (ICLR)*, New Orleans, LA, USA, 2019.

[159] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances in Neural Infor. Proc. Systems (NeurIPS)*, Long Beach, CA, USA, 2017, p. 1876–1887.

[160] G. Carbajal, J. Richter, and T. Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 681–685.

[161] Y. Du, K. Sekiguchi, Y. Bando, A. A. Nugraha, M. Fontaine, K. Yoshii, and T. Kawahara, "Semi-supervised multichannel speech separation based on a phone- and speaker-aware deep generative model of speech spectrograms," in *Proc. Euro. Signal Proc. Conf. (EUSIPCO)*, Dublin, Ireland, 2021, pp. 870–874.

[162] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural computation*, vol. 31, no. 9, pp. 1891–1914, 2019.

[163] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks." in *ISCA Interspeech*, Shanghai, China, 2020, pp. 4516–4520.

[164] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 371–375.

[165] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, "Dynamical variational autoencoders: A comprehensive review," *Foundations and Trends® in Machine Learning*, vol. 15, no. 1-2, pp. 1–175, 2021.

[166] J. Lian, C. Zhang, and D. Yu, "Robust disentangled variational speech representation learning for zero-shot voice conversion," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Singapore, Singapore, 2022, pp. 6572–6576.

[167] G. Carbajal, J. Richter, and T. Gerkmann, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement," in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, New Paltz, New York, USA, 2021, pp. 126–130.

[168] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoders," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 28, pp. 1788–1800, 2020.

[169] M. Sadeghi and X. Alameda-Pineda, "Mixture of inference networks for VAE-based audio-visual speech enhancement," *IEEE Trans. on Signal Processing*, vol. 69, pp. 1899–1909, 2021.

[170] V.-N. Nguyen, M. Sadeghi, E. Ricci, and X. Alameda-Pineda, "Deep variational generative models for audio-visual speech separation," in *IEEE Int. Workshop on Machine Learning for Signal Proc. (MLSP)*, Gold Coast, Australia, 2021, pp. 1–6.

[171] M. Sadeghi and X. Alameda-Pineda, "Robust unsupervised audio-visual speech enhancement using a mixture of variational autoencoders," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 7534–7538.

[172] ——, "Switching variational auto-encoders for noise-agnostic audio-visual speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 6663–6667.

[173] K. Sekiguchi, Y. Bando, K. Yoshii, and T. Kawahara, "Bayesian multichannel speech enhancement with a deep speech prior," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Honolulu, Hawaii, USA, 2018, pp. 1233–1239.

[174] M. Pariente, A. Deleforge, and E. Vincent, "A statistically principled and computationally efficient approach to speech enhancement using variational autoencoders," in *ISCA Interspeech*, Graz, Austria, 2019, pp. 3158–3162.

[175] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Computation*, vol. 31, no. 9, pp. 1891–1914, 2019.

[176] L. Girin, F. Roche, T. Hueber, and S. Leglaive, "Notes on the use of variational autoencoders for speech and audio spectrogram modeling," in *International Conference on Digital Audio Effects (DAFx)*, Birmingham, UK, 2019, pp. 1–8.

[177] X.-L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

[178] G. Ince, K. Nakadai, T. Rodemann, J.-i. Imura, K. Nakamura, and H. Nakajima, "Assessment of single-channel ego noise estimation methods," in *IEEE/RSJ International Conf. on Intelligent Robots and Systems (IROS)*, San Francisco, California, USA, 2011, pp. 106–111.

[179] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[180] R. D. Nathoo, M. Kegler, and M. Stamenovic, "Two-step knowledge distillation for tiny speech enhancement," in *ICASSP*, Seoul, Korea, 2024, pp. 10 141–10 145.

[181] S. Kim and M. Kim, "Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation," in *WASPAA*, Toronto, Canada, 2021, pp. 176–180.

[182] M. Sadeghi and R. Serizel, "Fast and efficient speech enhancement with variational autoencoders," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.

[183] P. Smaragdis and S. Venkataramani, "A neural network alternative to non-negative audio models," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, New Orleans, LA, USA, 2017, pp. 86–90.

[184] IEEE, "Information for Authors, IEEE Signal Processing Society," https://signalprocessingsociety.org/publications-resources/information-authors.

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, 2025

Huajian Fang

Appendix A

The journal publication [P2] extends the conference publication [P1] according to the policy stated in [184], so there is content overlap between them. [P1] is omitted in the main body for brevity and included here for completeness.

A.1 [P1] Integrating Statistical Uncertainty Into Neural Network-Based Speech Enhancement

Reference

H. Fang, T. Peer, S. Wermter, and T. Gerkmann, "Integrating statistical uncertainty into neural network-based speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Singapore, Singapore, 2022, pp. 386–390. DOI: 10.1109/ICASSP43922.2022.9747642

Copyright Notice

INTEGRATING STATISTICAL UNCERTAINTY INTO NEURAL NETWORK-BASED SPEECH ENHANCEMENT

Huajian Fang[1,2]*, Tal Peer*[1]*, Stefan Wermter*[2]*, Timo Gerkmann*[1]

[1]Signal Processing (SP), Universität Hamburg, Germany
[2]Knowledge Technology (WTM), Universität Hamburg, Germany
{huajian.fang,tal.peer,stefan.wermter,timo.gerkmann}@uni-hamburg.de

ABSTRACT

Speech enhancement in the time-frequency domain is often performed by estimating a multiplicative mask to extract clean speech. However, most neural network-based methods perform point estimation, i.e., their output consists of a single mask. In this paper, we study the benefits of modeling uncertainty in neural network-based speech enhancement. For this, our neural network is trained to map a noisy spectrogram to the Wiener filter and its associated variance, which quantifies uncertainty, based on the *maximum a posteriori* (MAP) inference of spectral coefficients. By estimating the distribution instead of the point estimate, one can model the uncertainty associated with each estimate. We further propose to use the estimated Wiener filter and its uncertainty to build an approximate MAP (A-MAP) estimator of spectral magnitudes, which in turn is combined with the MAP inference of spectral coefficients to form a hybrid loss function to jointly reinforce the estimation. Experimental results on different datasets show that the proposed method can not only capture the uncertainty associated with the estimated filters, but also yield a higher enhancement performance over comparable models that do not take uncertainty into account.

Index Terms— Speech enhancement, uncertainty estimation, Wiener filter, Bayesian estimator, deep neural network

1. INTRODUCTION

Single-channel speech enhancement algorithms typically operate in the short-time Fourier transform (STFT) domain [1]–[3]. The Gaussian statistical model in the STFT domain has been shown to be effective [1], [4]. Given the assumption that the complex-valued speech and noise coefficients are uncorrelated and Gaussian-distributed with zero mean, various estimators have been derived, such as the Wiener filter and the short-time spectral amplitude (STSA) estimator [1], [4], [5]. The Wiener filter, which is optimal in the minimum mean squared error (MMSE) sense, requires estimation of speech and noise variances. This can be achieved by various signal processing estimators with varying degrees of success for different signal characteristics [1], [2], [6]–[11].

Recently, deep neural networks (DNNs) have been successfully applied to speech enhancement and regularly show an improved performance over classical methods [10]–[13]. Among the DNN-based approaches relevant to this work are deep generative models (e.g., variational autoencoder) and supervised masking approaches. Generative models estimate the clean speech distribution and subsequently combine it with a separate noise model to construct a point estimate of a noise-removing mask (Wiener filter) [10], [11]. In contrast, typical supervised learning approaches are trained on pairs of noisy and clean speech samples and directly estimate a time-frequency mask that aims at reducing noise interference with minimal speech distortion given a noisy mixture, using a suitable loss function (e.g., mean squared error (MSE)) [12], [13]. However, the supervised approaches often learn the mapping between noisy and clean speech blindly and output a single point estimate without guarantee or measure of its accuracy. In this work we focus on adding an uncertainty measure to a supervised method by estimating the speech posterior distribution, instead of only its mean. Note that while this is conceptually related to the generative approach, in this case we do not estimate the clean speech prior distribution, but rather the posterior distribution of clean speech given a noisy mixture.

Uncertainty modeling based on neural networks has been actively studied in e.g., computer vision [14]. Inspired by this, here we propose a hybrid loss function to capture uncertainty associated with the estimated Wiener filter in the neural network-based speech enhancement algorithm, as depicted in Fig. 1. More specifically, we propose to train a neural network to predict the Wiener filter and its variance, which quantifies the uncertainty, based on the *maximum a posteriori* (MAP) inference of complex spectral coefficients, such that full Gaussian posterior distribution can be estimated. To regularize the variance estimation, we build an approximate MAP (A-MAP) estimator of spectral magnitudes using the estimated Wiener filter and uncertainty, which is in turn used together with the MAP inference of spectral coefficients to form a hybrid loss function. Experimental results show the effectiveness of the proposed approach in capturing uncertainty. Furthermore, the A-MAP estimator based on the estimated Wiener filter and its associated uncertainty results in improved speech enhancement performance.

2. SIGNAL MODEL

We consider the speech enhancement problem in the single microphone case with additive noise. The noisy signal x can be transformed into the time-frequency domain using the STFT:

$$X_{ft} = S_{ft} + N_{ft}, \tag{1}$$

where X_{ft}, S_{ft}, and N_{ft} are complex noisy speech coefficients, complex clean speech coefficients, and complex noise coefficients, respectively. The frequency and frame indices are given by $f \in \{1,2,\cdots,F\}$ and $t \in \{1,2,\cdots,T\}$, where F denotes the number of frequency bins, and T represents the number of time frames. Furthermore, we assume a Gaussian statistical model, where the speech and noise coefficients are uncorrelated and follow a circularly symmetric complex Gaussian distribution with zero mean, i.e.,

$$S_{ft} \sim \mathcal{N}_{\mathbb{C}}(0,\sigma^2_{s,ft}), \quad N_{ft} \sim \mathcal{N}_{\mathbb{C}}(0,\sigma^2_{n,ft}), \tag{2}$$

where $\sigma^2_{s,ft}$ and $\sigma^2_{n,ft}$ represent the variances of speech and noise, respectively. The likelihood $p(X_{ft}|S_{ft})$ follows a complex Gaussian

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 247465126

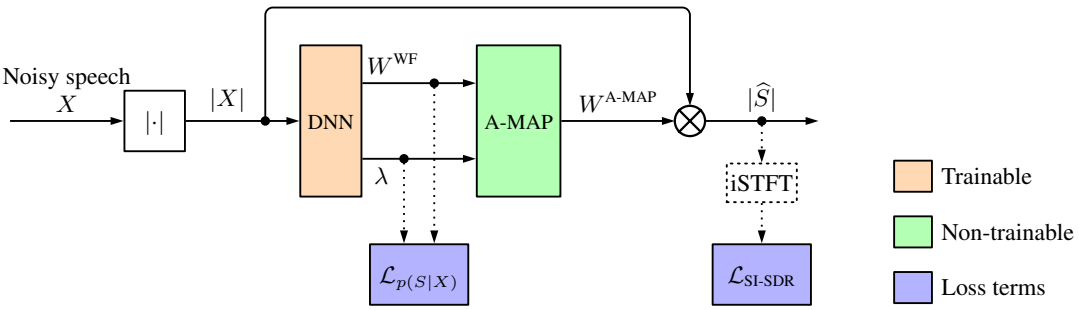


Fig. 1. Block diagram of the neural network-based uncertainty estimation. The neural network is trained according to the proposed hybrid loss function.

distribution with mean S_{ft} and variance $\sigma^2_{n,ft}$, given by

$$p(X_{ft}|S_{ft}) = \frac{1}{\pi\sigma^2_{n,ft}} \exp\left(-\frac{|X_{ft}-S_{ft}|^2}{\sigma^2_{n,ft}}\right). \tag{3}$$

Given the speech prior in (2) and the likelihood in (3), we can apply Bayes' theorem to find the speech posterior distribution, which is complex Gaussian of the form

$$p(S_{ft}|X_{ft}) = \frac{1}{\pi\lambda_{ft}} \exp\left(-\frac{|S_{ft}-W^{\text{WF}}_{ft}X_{ft}|^2}{\lambda_{ft}}\right), \tag{4}$$

where $W^{\text{WF}}_{ft} = \frac{\sigma^2_{s,ft}}{\sigma^2_{s,ft}+\sigma^2_{n,ft}}$ is the *Wiener filter* and $\lambda_{ft} = \frac{\sigma^2_{s,ft}\sigma^2_{n,ft}}{\sigma^2_{s,ft}+\sigma^2_{n,ft}}$ is the posterior's variance [1]. The MMSE and MAP estimators of S_{ft} under this model are both given by the *Wiener filter* [1]: $\widetilde{S}_{ft} = W^{\text{WF}}_{ft}X_{ft}$. It is known that the expectation of MMSE estimation error is closely related to the posterior variance [15], and under the assumption of complex Gaussian distribution it corresponds directly to the variance, i.e.,

$$\begin{aligned} E\{|S_{ft}-\widetilde{S}_{ft}|^2\} &= \iint |S_{ft}-\widetilde{S}_{ft}|^2 p(S_{ft}|X_{ft})p(X_{ft})\mathrm{d}S_{ft}\mathrm{d}X_{ft} \\ &= \int \lambda_{ft} p(X_{ft})\mathrm{d}X_{ft} = \lambda_{ft}. \end{aligned} \tag{5}$$

The variance λ_{ft} can be interpreted as a measure of uncertainty associated with the MMSE estimator [1]. In the following sections λ_{ft} will be referred to as the (estimation) uncertainty.

3. DEEP UNCERTAINTY ESTIMATION

The Wiener filter can be computed for a given noisy signal by estimation of $\sigma^2_{s,ft}$ and $\sigma^2_{n,ft}$ using traditional signal processing techniques. It is, however, also possible to directly estimate W^{WF}_{ft} using a DNN. Furthermore, if optimization is based on the posterior (4), besides W^{WF}_{ft} also the uncertainty λ_{ft} can be estimated as previously proposed in the computer vision domain [14]. Taking the negative logarithm (which does not affect the optimization problem due to monotonicity) and averaging over the time-frequency plane results in the following minimization problem:

$$\widetilde{W}^{\text{WF}}_{ft}, \widetilde{\lambda}_{ft} = \underset{W^{\text{WF}}_{ft},\lambda_{ft}}{\operatorname{argmin}} \underbrace{\frac{1}{FT}\sum_{f,t} \log(\lambda_{ft}) + \frac{|S_{ft}-W^{\text{WF}}_{ft}X_{ft}|^2}{\lambda_{ft}}}_{\mathcal{L}_{p(S|X)}}, \tag{6}$$

where $\widetilde{W}_{ft}$, $\widetilde{\lambda}_{ft}$ denote estimates of the Wiener filter and its uncertainty. If we assume a constant uncertainty for all time-frequency bins, i.e., $\lambda_{ft} = \lambda^*$, and refrain from explicitly optimizing for λ^*, $\mathcal{L}_{p(S|X)}$ degenerates into the well known MSE loss

$$\mathcal{L}_{\text{MSE}} = \frac{1}{FT}\sum_{f,t} |S_{ft}-W^{\text{WF}}_{ft}X_{ft}|^2, \tag{7}$$

which is widely used in DNN-based regression tasks, including speech enhancement [12], [16]. In this work we depart from the assumption of constant uncertainty. Instead, we propose to include uncertainty estimation as an additional task by training a DNN with the full negative log-posterior $\mathcal{L}_{p(S|X)}$.

It has been previously shown that modeling uncertainty by minimizing $\mathcal{L}_{p(S|X)}$ results in improvement over baselines that do not take uncertainty into account in computer vision tasks [14]. However, in preliminary experiments we have observed that directly using (6) as loss function results in reduced estimation performance for the Wiener filter and is prone to overfitting. To overcome this problem, we propose an additional regularization of the loss function by incorporating the estimated uncertainty into clean speech estimation as described next.

4. JOINT ENHANCEMENT AND UNCERTAINTY ESTIMATION

Besides estimation of the Wiener filter and its uncertainty, we propose to also incorporate a subsequent speech enhancement task that explicitly uses both into the training procedure. The speech enhancement task provides additional coupling between the DNN outputs (Wiener filter and uncertainty). In this manner, the DNN is guided towards estimation of uncertainty values that are relevant to the speech enhancement task, as well as enhanced estimation of the Wiener filter.

If we consider complex coefficients with symmetric posterior (4), the MAP and MMSE estimators both result directly in the Wiener filter W^{WF}_{ft} and do not require an uncertainty estimate. However, this changes if we consider spectral magnitude estimation. The magnitude posterior $p(|S_{ft}|\,|X_{ft})$, found by integrating the phase out of (4), follows a Rician distribution [5]

$$\begin{aligned} &p(|S_{ft}||X_{ft}) = \\ &\frac{2|S_{ft}|}{\lambda_{ft}} \exp\left(-\frac{|S_{ft}|^2+(W^{\text{WF}}_{ft})^2|X_{ft}|^2}{\lambda_{ft}}\right) I_0\left(\frac{2|X_{ft}||S_{ft}|W^{\text{WF}}_{ft}}{\lambda_{ft}}\right), \end{aligned} \tag{8}$$

where $I_0(\cdot)$ is the modified zeroth-order Bessel function of the first kind.

In order to compute the MAP estimate for the spectral magnitude, one needs to find the mode of the Rician distribution, which is difficult to do analytically. However, one may approximate it with a simple closed-form expression [5]:

$$|\widehat{S}_{ft}| \approx W_{ft}^{\text{A-MAP}} |X_{ft}| = \left(\frac{1}{2} W_{ft}^{\text{WF}} + \sqrt{\left(\frac{1}{2} W_{ft}^{\text{WF}} \right)^2 + \frac{\lambda_{ft}}{4|X_{ft}|^2}} \right) |X_{ft}|, \quad (9)$$

where $|\widehat{S}_{ft}|$ is an estimate of the clean spectral magnitude $|S_{ft}|$ using the A-MAP estimator of spectral magnitudes $W_{ft}^{\text{A-MAP}}$. It can be seen that the estimator $W_{ft}^{\text{A-MAP}}$ makes use of both the Wiener filter W_{ft}^{WF} and the associated uncertainty λ_{ft}. An estimate of the time-domain clean speech signal, denoted as $\widehat{s}$, is then obtained by combining the estimated magnitude $|\widehat{S}_{ft}|$ with the noisy phase, followed by the inverse STFT (iSTFT). The estimated time-domain signal is then used to compute the negative scale-invariant signal-to-distortion ratio (SI-SDR) metric [17]:

$$\mathcal{L}_{\text{SI-SDR}} = -10\log_{10}\left(\frac{||\alpha s||^2}{||\alpha s - \widehat{s}||^2} \right), \quad \alpha = \frac{\widehat{s}^T s}{||s||^2}, \quad (10)$$

which is in turn used as an additional term in the loss function that forces the speech estimate (computed with $W_{ft}^{\text{A-MAP}}$) to be similar to the clean target s.

Finally, we propose to combine the SI-SDR loss $\mathcal{L}_{\text{SI-SDR}}$ with the negative log-posterior $\mathcal{L}_{p(S|X)}$ given in (6), and train the neural network using a hybrid loss

$$\mathcal{L} = \beta \mathcal{L}_{p(S|X)} + (1-\beta)\mathcal{L}_{\text{SI-SDR}}, \quad (11)$$

with the weighting factor $\beta \in [0,1]$ as the hyperparameter. By explicitly using the estimated uncertainty for the speech enhancement task, the hybrid loss guides both mean and variance estimation to improve speech enhancement performance. An overview of this approach is depicted in Fig. 1.

5. EXPERIMENTAL SETTING

5.1. Dataset

For training we use the Deep Noise Suppression (DNS) Challenge dataset [18], which includes a large amount of synthesized noisy and clean speech pairs. We randomly sample a subset of 100 hours with signal-to-noise ratios (SNRs) uniformly distributed between -5 dB and 20 dB. The data are randomly split into training and validation sets (80% and 20% respectively).

Evaluation was performed on the synthetic test set without reverberation from DNS Challenge. Noisy signals are generated by mixing clean speech signals from [19] with noise clips sampled from 12 noise categories [18], with SNRs uniformly drawn from 0 dB to 25 dB. To examine performance across different datasets, we additionally synthesized another test dataset using clean speech signals from the `si_et_05` subset of the WSJ0 [20] dataset and four types of noise signals from CHiME [21] (`cafe`, `street`, `pedestrian`, and `bus`) with SNRs randomly sampled from {-10 dB, -5 dB, 0 dB, 5 dB, 10 dB}. A few samples are dropped due to the clipping effect in the mixing processing, and finally, this results in a test dataset of 623 files.

5.2. Baselines

To evaluate the effectiveness of modeling uncertainty in neural network-based speech enhancement, we consider training the same neural network using standard cost functions, i.e., the MSE defined as $\mathcal{L}_{\text{MSE}}$ in (7) and the SI-SDR defined as $\mathcal{L}_{\text{SI-SDR}}$ in (10). They are represented by MSE and SI-SDR in Table 1 and Fig. 3.

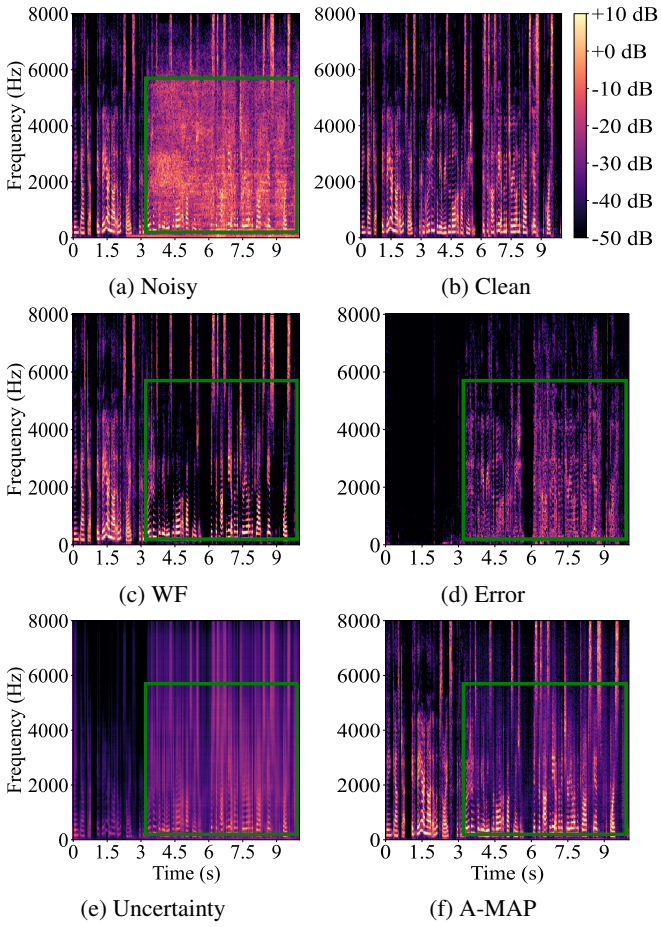


Fig. 2. Example of estimation uncertainty captured by the proposed method on the DNS test dataset, shown in (e). The proposed method allows estimating clean speech by either using the estimated Wiener filter or applying the A-MAP estimator that incorporates both the estimated Wiener filter and the associated uncertainty, and the resulting estimates are shown in (c) and (f), denoted by WF and A-MAP, respectively. The estimation error of Wiener filtering in (d) is computed between the estimated magnitudes (c) and clean magnitudes (b), indicating over- or under-estimation of speech magnitudes.

5.3. Hyperparameters

All audio signals are sampled at 16 kHz and transformed into the time-frequency domain using the STFT with a 32 ms Hann window and 50% overlap.

For a fair comparison, we used the separator of Conv-TasNet [22] that has a temporal convolution network (TCN) architecture. It has been shown to be effective in modeling temporal correlations. We used the causal version of the implementation and default hyperparameters provided by the authors[1] without performing a hyperparameter search. Note that for our model performing uncertainty estimation, the output layer is split into two heads that predict both the Wiener filter and the uncertainty. We applied the sigmoid activation function to the estimated mask, while using the *log-exp* technique to constrain the uncertainty output to be greater than 0, i.e., the network outputs the logarithm of the variance, which is then recovered by the exponential term in the loss function. All neural networks were trained for 50 epochs with a batch

[1] `https://github.com/naplab/Conv-TasNet`

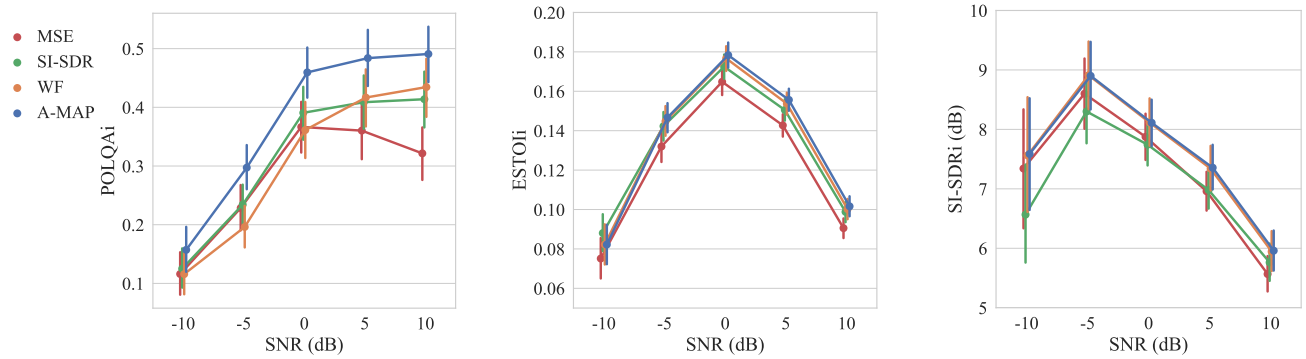


Fig. 3. Performance improvement obtained on the synthetic dataset using clean speech from WSJ0 and noise signals from CHiME. POLQAi denotes POLQA improvement relative to noisy mixtures. The same definition applies to ESTOIi and SI-SDRi. The marker denotes the mean value over all utterances and the vertical bar indicates the 95%-confidence interval.

	POLQA	ESTOI	SI-SDR (dB)
Noisy	2.30 ± 0.10	0.81 ± 0.02	9.07 ± 0.89
SI-SDR	2.93 ± 0.11	0.88 ± 0.01	15.99 ± 0.75
MSE	2.88 ± 0.10	0.88 ± 0.01	16.05 ± 0.71
Proposed WF	3.00 ± 0.11	0.88 ± 0.01	16.39 ± 0.73
Proposed A-MAP	$\mathbf{3.06 \pm 0.10}$	$\mathbf{0.89 \pm 0.01}$	$\mathbf{16.42 \pm 0.73}$

Table 1. Average performance over all utterances of the DNS non-reverberant synthetic test dataset in terms of POLQA, ESTOI, and SI-SDR. Values are given in mean $\pm$ confidence interval (95% confidence).

size 16, the maximum norm of gradients was set to 5, and the parameters were optimized using the Adam optimizer [23] with a learning rate of 0.001. We halved the learning rate if the validation loss did not decrease for 3 consecutive epochs. To prevent overfitting, training was stopped if the validation loss failed to decrease within 10 consecutive epochs. The weighting factor β is set to 0.01, chosen empirically.

6. RESULTS AND DISCUSSION

6.1. Analysis of uncertainty estimation

In Fig. 2, we use an audio example from the DNS test dataset to illustrate the uncertainty captured by the proposed method, and all plots are shown in decibel (dB) scale. Applying the estimated Wiener filter to the noisy coefficients yields an estimate of the clean speech, denoted as WF shown in Fig. 2 (c). To measure the prediction error, we can compute the absolute values of the difference between the estimated magnitudes, i.e., WF, and reference magnitudes given in Fig. 2 (b), which indicates over- or under-estimation of speech magnitudes, shown in Fig. 2 (d). It is observed that the model produces large errors when speech is heavily corrupted by noise, as can be seen by comparing the marking regions (green boxes) of the noisy mixture shown in Fig. 2 (a) and the prediction error of Fig. 2 (d). By comparing error in Fig. 2 (d) and uncertainty in Fig. 2 (e), the estimator generally associates large uncertainty with large prediction errors, while giving low uncertainty to accurate estimates, e.g., the first 3 seconds. This shows that the model produces uncertainty measurements that are closely related to estimation errors. In our proposed method with uncertainty estimation, we can use not only the estimated Wiener filter, but also the estimated A-MAP mask that incorporates both the estimated uncertainty and Wiener filter, as given in (9). This estimate is denoted as A-MAP in Fig. 2 (f). We observe that the A-MAP estimate causes less speech distortion compared with the WF estimate, as can be seen, e.g., from the marking regions of WF and A-MAP.

6.2. Performance Evaluation

In Table 1, we present average evaluation results of our method on the DNS synthetic test set in terms of SI-SDR measured in dB, extended short-time objective intelligibility (ESTOI) [24], and perceptual objective listening quality analysis (POLQA)[2] [25]. We observe that modeling uncertainty yields improvement over the baselines, where the proposed WF outperforms the baselines in terms of POLQA and SI-SDR, and a larger improvement can be observed between the baselines and the proposed A-MAP. This shows that it is advantageous to model uncertainty within the model instead of directly estimating optimal points.

In Fig. 3, we present speech enhancement results in terms of mean improvement of POLQA, ESTOI, and SI-SDR. For this evaluation we used another unseen test dataset based on speech from WSJ0 and noise from CHiME. It shows that our proposed approach performs better in terms of speech quality given by higher POLQA values without deteriorating ESTOI (with an exception at SNR of -10 dB) and SI-SDR, which again demonstrates the benefit of modeling uncertainty. We also observe that larger improvement over the baselines is achieved at high SNRs, which may be explained by the fact that, at high SNRs, speech quality (and thus POLQA) is mainly affected by speech distortions, while at low SNRs the main factor is residual noise.

7. CONCLUSION

Based on the common complex Gaussian model of speech and noise signals, we proposed to augment the existing neural network architecture with an additional uncertainty estimation task. Specifically, we proposed simultaneous estimation of the Wiener filter and the associated uncertainty to capture the full speech posterior distribution. Furthermore, we proposed using the estimated Wiener filter and uncertainty to produce an A-MAP estimate of the clean spectral magnitude. Eventually, we combined uncertainty estimation and speech enhancement by the proposed hybrid loss function. We showed that the approach can capture uncertainty and lead to improved speech enhancement performance across different speech and noise datasets. For future work, it would be interesting to integrate the uncertainty estimation into multi-modal learning systems, which may rely more on other modalities when audio modality raises high uncertainty.

[2]We would like to thank J. Berger and Rohde&Schwarz SwissQual AG for their support with POLQA.

8. REFERENCES

[1] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds., Wiley, 2018, pp. 65–85.

[2] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement: a survey of the state of the art*. Morgan & Claypool Publishers, 2013, pp. 1–80.

[3] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[5] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the ephraim and malah suppression rule for audio signal enhancement," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 10, pp. 1–9, 2003.

[6] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 145–148.

[7] G. Carbajal, J. Richter, and T. Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 681–685.

[8] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 676–680.

[9] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks.," in *INTERSPEECH*, 2020, pp. 4516–4520.

[10] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *International Workshop on Machine Learning for Signal Processing (MLSP)*, 2018, pp. 1–6.

[11] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 716–720.

[12] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[13] R. Rehr and T. Gerkmann, "SNR-based features and diverse training data for robust DNN-based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1937–1949, 2021.

[14] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[15] R. F. Astudillo, D. Kolossa, and R. Orglmeister, "Accounting for the uncertainty of speech estimates in the complex domain for minimum mean square error speech enhancement," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[16] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *44th International Conference on Telecommunications and Signal Processing (TSP)*, 2021, pp. 72–76.

[17] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 626–630.

[18] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, *et al.*, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *arXiv preprint arXiv:2005.13981*, 2020.

[19] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[20] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Sennheiser LDC93S6B," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[21] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.

[22] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, Dec. 2014.

[24] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[25] ITU-T Rec. P.863, *Perceptual objective listening quality prediction*, *International Telecommunication Union*, Geneva, Switzerland, 2011.

Appendix B

B.1 Derivation of the Update Rule

The iterative update rules in [P7] are derived by taking the derivative of the auxiliary function with respect to the relevant parameters. The derivation of the update rule for the activation matrix is detailed in the accompanying support document, which is linked in the paper and openly available. It is included below for completeness.

Supporting document to the paper "Joint Reduction of Ego-noise and Environmental Noise with a Partially-adaptive Dictionary"

Huajian Fang, Guillaume Carbajal, Stefan Wermter, Timo Gerkmann

This supporting document provides the derivation of the update for the parameter H as an example. Recall that given the VAE speech model and two NMF noise models, the noisy mixture can be described by

$$x_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, g_t \hat{\sigma}^2_{s,f}(z_t) + \sum_k w_{fk} h_{kt} + \sum_l \overline{w}_{fl} \overline{h}_{lt}) \tag{1}$$

It is intractable to compute the maximum likelihood of the model with latent variable z_t and unknown parameters $\zeta = \{w_{fk}, h_{kt}, \overline{w}_{fl}, \overline{h}_{lt}, g_t\}$ directly. However, it can be alternatively solved by the Monte Carlo expectation maximization (MCEM) algorithm [2, 3].

The expectation of the complete data log-likelihood is shown as

$$\begin{aligned} Q(\zeta, \zeta^\sharp) &= E_{p(Z|X,\zeta^\sharp)}[\ln p(X, Z|\zeta)] \\ &\simeq -\frac{1}{R} \sum_r \left(\sum_{f,t} \left(\ln \left(g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + \sum_k w_{fk} h_{kt} + \sum_l \overline{w}_{fl} \overline{h}_{lt} \right) \right. \right. \\ &\left. \left. + \frac{|x_{ft}|^2}{g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + \sum_k w_{fk} h_{kt} + \sum_l \overline{w}_{fl} \overline{h}_{lt}} \right) + TF \ln(\pi) \right), \end{aligned} \tag{2}$$

where the last constant term can be ignored and $\zeta^\sharp$ denotes an initialization of the parameters.

The unknown parameters are optimized using the auxiliary function technique [1]. Let $C(H)$ be defined as $-Q(\zeta, \zeta^\sharp)$ with respect to the parameter H only, whereas the other parameters are seen as constants. The auxiliary function $\mathcal{G}(H|\widetilde{H}) : \mathbb{R}_+^{K\times T} \times \mathbb{R}_+^{K\times T} \to \mathbb{R}_+$ is defined as the upper bound of $C(H)$ which is tight at $\widetilde{H}$ if and only if

- $\forall H \in \mathbb{R}_+^{K\times T}, C(H) = \mathcal{G}(H|H)$
- $\forall (H, \widetilde{H}) \in \mathbb{R}_+^{K\times T} \times \mathbb{R}_+^{K\times T}, C(H) \leq \mathcal{G}(H|\widetilde{H})$.

If the $(i+1)$-th iteration $H^{(i+1)}$ satisfies $\mathcal{G}(H^{(i+1)}|H^{(i)}) \leq \mathcal{G}(H^{(i)}|H^{(i)})$, $C(H^{(i+1)})$ and $C(H^{(i)})$ will meet

$$C(H^{(i+1)}) \leq \mathcal{G}(H^{(i+1)}|H^{(i)}) \leq \mathcal{G}(H^{(i)}|H^{(i)}) = C(H^{(i)}). \tag{3}$$

$H^{(i+1)}$ is chosen by minimizing:

$$H^{i+1} = \underset{H \in \mathbb{R}_+^{K\times T}}{\arg\min}\ \mathcal{G}(H|H^{(i)}). \tag{4}$$

This turns optimization of the criterion function $C(H)$ into iterative optimization of the auxiliary function.

By decomposing $C(H)$ into concave and convex parts, denoted by $\hat{C}(H)$ and $\breve{C}(H)$, respectively, auxiliary functions can be constructed for each part separately [1]. The decomposition is shown as

$$C(H) = \breve{C}(H) + \hat{C}(H) \tag{5}$$

with

$$\hat{C}(H) = \frac{1}{R} \sum_r \sum_{f,t} \left(\ln \left(g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + (WH)_{ft} + \left(\overline{W}\overline{H}\right)_{ft} \right) \right), \tag{6}$$

$$\breve{C}(H) = \frac{1}{R} \sum_r \sum_{f,t} \left(\frac{|x_{ft}|^2}{g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + (WH)_{ft} + \left(\overline{W}\overline{H}\right)_{ft}} \right), \tag{7}$$

where $(WH)_{ft} = \sum_k w_{fk} h_{kt}$ and $\left(\overline{W}\overline{H}\right)_{ft} = \sum_l \overline{w}_{fl} \overline{h}_{lt}$.

- **Concave:**

An auxiliary function $\hat{\mathcal{G}}(H, \widetilde{H})$ to the concave part $\hat{C}(H)$ can be defined as its first order Taylor approximation at $\widetilde{H}$ using the upper-bound property of its tangent:

$$\hat{C}(H) \leq \hat{\mathcal{G}}(H, \widetilde{H}) = \hat{C}(\widetilde{H}) + \nabla^T \hat{C}(\widetilde{H})(H - \widetilde{H}) \tag{8}$$

This gives:

$$\hat{\mathcal{G}}(H, \widetilde{H}) = \frac{1}{R} \sum_r \sum_{f,t} \left(\ln \left(g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + \left(W\widetilde{H}\right)_{ft} + \left(\overline{W}\overline{H}\right)_{ft} \right) + \frac{(WH)_{ft} - \left(W\widetilde{H}\right)_{ft}}{g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + \left(W\widetilde{H}\right)_{ft} + \left(\overline{W}\overline{H}\right)_{ft}} \right), \tag{9}$$

where $\left(W\widetilde{H}\right)_{ft} = \sum_k w_{fk}\widetilde{h}_{kt}$.

- **Convex**

An auxiliary function $\breve{\mathcal{G}}(H, \widetilde{H})$ to the convex part $\breve{C}(H)$ can be obtained using Jensen's inequality. The convex part can be written as:

$$\breve{C}(H) = \frac{1}{R} \sum_r \sum_{f,t} |x_{ft}|^2 \breve{C}^{(r)}_{ft}(h_t) \tag{10}$$

with

$$\breve{C}^{(r)}_{ft}(h_t) = \frac{1}{g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + (WH)_{ft} + \left(\overline{W}\overline{H}\right)_{ft}}, \tag{11}$$

where $h_t \in \mathbb{R}^K$ is the t-th column in H.

To construct Jensen's formula, we can define two terms:

$$a = \left(\frac{g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + \left(\overline{W}\overline{H}\right)_{ft}}{g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + \left(W\widetilde{H}\right)_{ft} + \left(\overline{W}\overline{H}\right)_{ft}} \right), \tag{12}$$

$$b = \left(\frac{\sum_k w_{fk}\widetilde{h}_{kt}}{g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + \left(W\widetilde{H}\right)_{ft} + \left(\overline{W}\overline{H}\right)_{ft}} \right), \tag{13}$$

such that

$$a + b = 1. \tag{14}$$

Applying Jensen's inequality to $\breve{C}^{(r)}_{ft}(h_t)$ gives:

$$\breve{C}^{(r)}_{ft}(h_t) = \frac{1}{a\frac{\left(g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + \left(\overline{W}\overline{H}\right)_{ft}\right)}{a} + b\frac{\left(\sum_k w_{fk}h_{kt}\right)}{b}} \leq a\frac{a}{g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + \left(\overline{W}\overline{H}\right)_{ft}} + b\frac{b}{\sum_k w_{fk}h_{kt}} = \breve{\mathcal{G}}_{ft}(h_t, \widetilde{h}_t) \tag{15}$$

From (10), (12), (13), and (15), $\breve{\mathcal{G}}(H, \widetilde{H})$ is given as:

$$\breve{\mathcal{G}}(H, \widetilde{H}) = \frac{1}{R} \sum_r \sum_{f,t} |x_{ft}|^2 \left(\frac{g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + \left(\overline{W}\overline{H}\right)_{ft}}{\left(g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + \left(W\widetilde{H}\right)_{ft} + \left(\overline{W}\overline{H}\right)_{ft} \right)^2} + \sum_k \frac{w_{fk}\widetilde{h}^2_{kt}}{h_{kt}\left(g_t \hat{\sigma}^2_{s,f}(z_t^{(r)}) + \left(W\widetilde{H}\right)_{ft} + \left(\overline{W}\overline{H}\right)_{ft} \right)^2} \right) \tag{16}$$

Finally, the complete auxiliary function is given by:

$$\mathcal{G}(H, \widetilde{H}) = \hat{\mathcal{G}}(H, \widetilde{H}) + \breve{\mathcal{G}}(H, \widetilde{H}) \tag{17}$$

Taking the derivative of the auxiliary function with respect to H and setting to zero will give the update rule. The optimization starts with random initialization, and the value of $\tilde{h}_{kt}$ is set to h_{kt} at the previous step. The derivation for other parameters can be done similarly.

References

[1] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, 2011.

[2] S. Leglaive, L. Girin, and R. Horaud. A variance modeling framework based on variational autoencoders for speech enhancement. In *International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2018.

[3] Greg CG Wei and Martin A Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.