DISSERTATION

Model-based Techniques and Diffusion Models for Speech Dereverberation

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

eingereicht von

Jean-Marie Lemercier

Hamburg 2024

Jean-Marie Lemercier: Model-based Techniques and Diffusion Models for Speech Dereverberation

GUTACHTER:

Prof. Dr.-Ing. Timo Gerkmann

Prof. Dr. Simon Doclo

Dr. Antoine Deleforge

VORSITZ DER PRÜFUNGSKOMMISSION:

Prof. Dr. Frank Steinicke

TAG DER EINREICHUNG:

04.11.2024

TAG DER DISPUTATION:

14.02.2025

Zusammenfassung

Nachhall tritt in vielen Umgebungen auf und beeinträchtigt oft die Verständlichkeit und Qualität der menschlichen Sprache. Die Auswirkungen auf Hörgeschädigte sind dabei noch gravierender. Die Entwicklung von Technologien für Multimedia-Unterhaltung, Kommunikation und medizinische Anwendungen hat inzwischen zu einer steigenden Nachfrage nach verbesserter Klangqualität geführt. Daher enthalten viele eingebettete Geräte heutzutage einen Enthallungsalgorithmus, der darauf abzielt, die anechoische Komponenten der Sprache wiederherzustellen. Enhalltung ist eine anspruchsvolle Aufgabe und ein schlecht gestelltes inverses Problem: Selbst bei perfekter Kenntnis der Raumakustik ist nicht garantiert, dass ein völlig enthalltes Signal erzeugt kann. Darüber hinaus ist in den meisten praktischen Fällen ein solches Wissen nicht verfügbar, weshalb die meisten Enthallungsalgorithmen als blind bezeichnet werden. Das bedeutet, dass sie alle Informationen aus dem verhallten Sprachsignal extrahieren müssen.

Traditionelle Enthallungsalgorithmen leiten anechoische Sprachschätzer her, indem sie statistische Eigenschaften von Sprachsignalen und deren Verteilung annehmen und, sofern verfügbar, Wissen über die Raumakustik ausnutzen. Herkömmliche Methoden funktionieren gut in ruhigen Umgebungen mit geringem Nachhall und wenig Hintergrundgeräuschen. Sie versagen jedoch, wenn die Bedingungen schwieriger werden oder die Annahmen, auf denen ihre Herleitung basiert, nicht erfüllt sind. Angesichts der jüngsten Fortschritte im datengesteuerten Deep Learning basieren viele aktuelle Algorithmen zur Enthallung von Sprache auf den Modellierungsfähigkeiten tiefer neuronaler Netze (DNNs). Diese leistungsstarken, nichtlinearen Schätzer ermöglichen es lernbasierten Ansätzen, traditionelle Ansätze bei komplexen Aufgaben wie der einkanaligen blinden Enthallung mit instationärem Messrauschen deutlich zu übertreffen. DNN-basierte Algorithmen benötigen jedoch mehr Rechenressourcen und leiden oft unter einer schlechten Anpassungsfähigkeit an Bedingungen, die in ihren Trainingsdaten nicht vorkommen, was zu anderen Fehlern führen kann als bei traditionellen Ansätzen. Darüber hinaus birgt die ausschließliche Nutzung DNN-basierter Lernansätze das Risiko, die Interpretierbarkeit zu verringern. Daher können DNN-basierte Ansätze zur Enthallung keine Garantien für die Sicherheit und Fairness der Nutzer bieten.

Das Eröffnungskapitel dieser Arbeit konzentriert sich auf modellbasiertes Lernen, d.h. auf hybride Paradigmen, die DNNs mit Domänenwissen kombinieren, wie z.B. statistische Eigenschaften von Sprache, Raumakustik oder Strukturen traditioneller Algorithmen. In der ersten Veröffentlichung stellen wir einen echtzeitfähigen, zweistufigen Algorithmus vor, der traditionelle Methoden der Sprachenthallung mit niedrig parametrisierten DNNs kombiniert. In der ersten Stufe entfernt ein mehrkanaliges lineares Vorhersageverfahren, das durch die Verwendung von DNNs unterstützt wird, den größten Teil des moderaten Nachhalls, der im Rahmen der linearen Vorhersagemöglichkeiten liegt. In der zweiten Stufe wird dann das finale Sprachsignal extrahiert, indem der statistisch unkorrelierte Restnachhall aus dem Ausgang der ersten Stufe unterdrückt wird. Die andere in diesem Kapitel vorgestellte Methode basiert

auf Signalmodellen, die der Sprachverbesserung und der Enthallung zugrunde liegen. Hier erweitern wir DNN-basierte Maskierungsansätze im Zeit-Frequenz-Bereich zu Ansätzen, die eine Multi-Frame-Filterung in Frequenz-Subbändern durchführen. Wir haben festgestellt, dass eine DNN-basierte Multi-Frame-Filterung bei der Enthallung besser abschneidet als die Maskierung einzelner Frames. Dies entspricht der intuitiven Erwartung, die sich aus den Konzepten der Subband-Filterung zur Enthallung ableiten lässt. Im Gegensatz dazu ist die Leistung beider Ansätze ähnlich, wenn nur Hintergrundgeräusche vorhanden sind.

Im zweiten Kapitel untersuchen wir bedingte, diffusionsbasierte generative Modelle zur Sprachenthallung und deren Beziehung zu überwachten Lernmethoden und prädiktiven Ansätzen. Bedingte generative Modelle schätzen die A-posteriori-Verteilung der anechoischen Sprache, gegeben einer verhallten Aufnahme. Im Gegensatz dazu lernen prädiktive Modelle eine direkte Abbildung zwischen verhallter und anechoischer Sprache. Wir leiten dieses Kapitel mit einem Tutorium über bedingte Diffusionsmodelle für die Audiorestauration ein. Der zweite Beitrag ist eine vergleichende Analyse von prädiktiven Methoden und diffusionsbasierten generativen Modellen. Wir analysieren diesen Vergleich im Kontext verschiedener Aufgaben der Sprachrestauration, wie Entrauschung, Enthallung und Bandbreitenerweiterung. Die Studie zeigt, dass Diffusionsmodelle ihre prädiktiven Gegenstücke bei allen Aufgaben konsequent übertreffen und dass der Qualitätsunterschied insbesondere bei nicht-additiven Störungstypen wie Nachhall und Bandbreitenerweiterung größer ist. Unsere letzte Veröffentlichung nutzt diese Analyse, um prädiktive Ansätze und diffusionsbasierte generative Modellierung auf prinzipielle Art und Weise zu kombinieren. Wir zeigen, dass die Verwendung einer prädiktiven Modellschätzung als Zwischenschritt vor der diffusionsbasierten Generierung zu einer erheblichen Verbesserung der Sprachqualität führt und gleichzeitig die Rechenkosten im Vergleich zur herkömmlichen Diffusionsmodellierung verringert.

Die Veröffentlichungen im letzten Kapitel dieser Dissertation behandeln die Enthallung als ein inverses Problem. Unser erster Beitrag stellt eine unüberwachte Lernmethode zur informierten Enthallung vor, bei der Diffusionsmodelle als A-priori-Wahrscheinlichkeit für saubere Sprache im Bayes'schen Posterior-Sampling eingesetzt werden. Wir stellen fest, dass die auf Diffusionsmodellen basierende A-priori-Wahrscheinlichkeit ein effektiver Regularisierer für die Lösung des inversen Problems ist, und eine gute Enthallung liefert, wenn die Raumakustik perfekt bekannt ist. Die zweite Arbeit erweitert die erste auf das nicht-informierte, blinde Szenario, bei dem die Raumakustik unbekannt ist. Basierend auf statistischen Beobachtungen der Raumeigenschaften schlagen wir vor, die Raumimpulsantwort mithilfe eines Subband-Filters mit frequenzabhängigen exponentiellem Abklang darzustellen. Der daraus resultierende Ansatz führt eine gemeinsame Enthallung und eine Schätzung der Raumimpulsantwort ohne jegliche Überwachung während des Trainings durch. Er zeichnet sich durch eine natürliche Anpassungsfähigkeit an neue, schallharte Umgebungen aus, da er unüberwacht trainiert wird. Dies unterscheidet ihn von überwachten Algorithmen, deren Leistung nachlässt, wenn sich die akustischen Bedingungen zur Testzeit von denen während des Trainings unterscheiden.

Zusammenfassend führt diese Dissertation eine gründliche Untersuchung der DNN-unterstützten Sprachenthallung durch, die von modellbasierten Techniken bis hin zu den neuesten Fortschritten in der diffusionsbasierten generativen Modellierung reicht. Wir diskutieren die Anwendbarkeit der in dieser Arbeit vorgestellten Methoden auf reale Anwendungen, wobei ein besonderer Schwerpunkt auf Hörgeräten liegt. Durch die verschiedenen in dieser Arbeit durchgeführten Studien können wir nachweisen, dass die Einbindung von Domänenwissen in DNN-basierte Verfahren entscheidend dazu beiträgt, interpretierbare und effiziente Algorithmen zur Enthallung von Sprache zu entwickeln.

Abstract

Reverberation occurs in most of our environments and often degrades the intelligibility and quality of human speech, with an aggravated effect on hearing-impaired listeners. Meanwhile, the evolution of technologies for multimedia entertainment, communications and medical applications has led to a greater demand for improved sound quality. Therefore, many embedded devices now include a dereverberation algorithm, which aims to recover the anechoic component of speech. Dereverberation is an arduous task and an ill-posed inverse problem: even perfectly knowing the room acoustics does not guarantee to obtain a perfectly dereverberated signal. Furthermore, in most real-life cases, such knowledge is not available and therefore most dereverberation algorithms are blind, i.e. they must extract information from the reverberant speech signal only.

Traditional dereverberation algorithms derive anechoic speech estimators exploiting statistical properties of speech signals, distributional assumptions and even knowledge of room acoustics when available. Traditional methods are efficient in quiet environments where reverberation and background noise are mild, but fail to perform satisfyingly when conditions become more adverse or when assumptions underlying their derivations do not hold. Given the recent shift toward data-driven deep learning, numerous speech dereverberation algorithms now rely on the impressive modelling capabilities of deep neural networks (DNNs). These powerful non-linear estimators allow learning-based approaches to largely outperform their traditional counterparts on tasks as difficult as single-channel blind speech dereverberation in the presence of non-stationary measurement noise. However, DNN-based algorithms require more computing resources and often suffer from poor adaptability to conditions unseen in their training data, leading to different failure cases than traditional techniques. Furthermore, relying solely on DNN-based learning approaches carries the risk of reducing interpretability, thus failing to provide guarantees with respect to user safety and fairness.

The opening chapter of this thesis focuses on model-based learning, i.e. hybrid paradigms combining DNNs with domain knowledge such as speech statistical properties, room acoustics or traditional algorithm structures. In the first publication, we present a real-time capable two-stage algorithm combining traditional speech dereverberation and lightweight DNNs. In the initial stage, a DNN-assisted multi-channel linear prediction method removes most of the moderate reverberation accessible within the auto-regressive filter length. The second stage then extracts the target speech by suppressing the statistically uncorrelated residual reverberation from the output of the first stage. The other technique presented in this chapter leverages the signal models behind speech denoising and dereverberation. There, we extend time-frequency masking DNNs to deep filters performing multi-frame filtering in frequency subbands. We observe that deep filters perform better on dereverberation than single-frame masking, as one would intuitively expect from the ideas underlying subband filtering for dereverberation. In contrast, the performance of both approaches is similar when only background noise is present.

In the second chapter, we investigate conditional diffusion-based generative models for speech dereverberation, and their relationship to supervised learning and predictive models. Conditional generative models estimate the posterior distribution of anechoic speech given a reverberant recording, in contrast with predictive models that learn a regression rule between reverberant and anechoic speech. We introduce this chapter with a tutorial on conditional diffusion models for audio restoration. The second contribution is a comparative analysis of predictive methods versus diffusion-based generative models. We contextualize this comparison with respect to various speech restoration tasks such as denoising, dereverberation and bandwidth extension. The study suggests that diffusion models consistently outperform their predictive counterparts across all tasks, and that the quality difference is larger for non-additive degradation models such as reverberation and bandwidth extension. Our next work leverages this analysis to combine predictive and diffusion-based generative models in a principled fashion. We demonstrate that using a predictive model estimate as an intermediate step before diffusion-based generation yields remarkable speech enhancement and dereverberation performance, while simultaneously reducing computational costs compared to traditional diffusion models.

The publications in the chapter 5 of this dissertation treat dereverberation as an inverse problem. Our initial contribution presents an unsupervised method for informed dereverberation, where diffusion models are applied as unconditional speech priors in Bayesian posterior sampling. We observe that the diffusion-based prior is an effective regularizer for inverse problem solving, yielding state-of-the-art dereverberation performance when the room acoustics are perfectly known. The second work extends the former to the blind scenario where room acoustics are unknown. Rooted in statistical observations of room properties, we propose to represent the room impulse response by a subband filter with frequency-dependent exponential decays. The resulting approach performs joint dereverberation and room impulse response estimation without any supervision during training. It boasts a natural adaptability to new reverberant environments because of the lack of supervision at training time, unlike supervised algorithms whose performance dwindles when acoustic conditions at test time are different from those seen during training.

In conclusion, this dissertation conducts a principled investigation of DNN-assisted speech dereverberation, ranging from model-based techniques to recent advances in diffusion-based generative models. We continually discuss the applicability of the methods presented in this thesis to real-life applications, with a particular focus on hearing devices. Through the various analyses run in this thesis, we provide evidence that injecting domain knowledge in DNN-based techniques is instrumental in providing interpretable and efficient speech dereverberation algorithms.

Acknowledgements

My deepest thanks and admiration go to Timo Gerkmann for his support and understanding over these years. Your team management skills are unrivaled and I could not have dreamed of a better supervisor during this thesis. Discussing with you has taught me a great lot about scientific honesty, rigor, and eventually what makes great research. Having such a kind and sincerely enthusiastic mentor is a luck, and I am particularly grateful for our personal relationship.

Whether research works or not, colleagues are the people that keep you going. That is why I want to deeply thank the SP team (from A to Z): Alina, Danilo, David, Guillaume, Hector, Huajian, Jakob, Julius, Kristina, Long, Navin, Simon and Tal. Some I got to share more personal moments with, some less, but all I could trust for everything. A massive shout out to our computer wizard Reinhard, whom we will sorely miss at the Informatikum. Good luck for everything and thanks for sorting out the continuous torture we inflicted to our poor machines. Thank you Stephanie for all the help with administrative matters and the kind supervision.

Talking about colleagues, I can not thank enough Eloi Moliner from Aalto University. I don't believe I have had such a fluid and fruitful collaboration with anyone before. I truely believe you are a gifted scientist and deserve the best this field can offer, all my wishes to you. Massive thanks to Simon Rouard for helping me turn my agitated internship at Meta into an ICML paper, plus being a great friend and colleague to be around.

All my respect and appreciation go to Alexandre Défossez for supervising my work at Meta, I learnt so much to your contact and really loved our mathy discussions. I don't think it is an overstatement to say that you are set for much more great research, so keep it coming.

My unconditional love goes to my family and friends: they are a precious luxury and it should never be forgotten.

This dissertation is also dedicated to the most valuable person in my life, my girlfriend and life accomplice Adé. Thank you for absolutely everything.

Table of Contents

Notations

In chapters 1 and 6, we will introduce each variable with its dimensions and arguments. We generally consider digital signals with discrete indexes, using the following notational conventions unless indicated otherwise:

- $s(n) \in \mathbb{R}$ (lowercase normal font) is a one-dimensional waveform, with time index n.
- $S(k, f) \in \mathbb{C}$ (uppercase normal font) is a one-dimensional time-frequency complex spectrogram obtained via the short-time Fourier transform (STFT), with time-frame index k and frequency bin f.
- $\mathbf{s}(n) := [s_1(n), \ldots, s_{D_1}(n)]^T \in \mathbb{R}^{D_1}$ (lowercase bold font) is a vector-valued waveform, such as, e.g. a multi-channel signal recorded at a microphone array.
- $\mathbf{S}(k, f) := [S_1(k, f), \ldots, S_{D_1}(k, f)]^T \in \mathbb{C}^{D_1}$ (uppercase bold font) is a vector-valued time-frequency complex spectrogram.
- $\mathcal{S} \in \mathbb{C}^{D_1 \times D_2 \times \cdots \times D_N}$ (calligraphied) is a second- or higher-order tensor, with potential arguments not shown here.

We advise the reader that these notations may slightly differ from those used in the publications composing this thesis in chapters 3, 4 and 5. This is because the publications were written prior to this dissertation and each of them in a potentially different mathematical context.

Own publications included by the author in this dissertation are referred to as [Pxx] in the bibliography, while those from other authors (or not included explicitly in this dissertation) follow the regular notation [xx].

1

Introduction

1.1 Motivation

Most environments in our everyday lives are reverberant. Indeed, room walls and other physical objects reflect the sound waves originating from sound sources in the room. Therefore, sound receivers, such as microphones or human ears, pick up not only the direct sound but also these many reflections resulting in the phenomenon of *reverberation*. This scenario is an instance of the *cocktail party problem* illustrated on Figure 1.1, which denotes an environment where reverberation, multiple competing speakers and noise sources are present. Normal-hearing listeners are particularly robust to most reverberant conditions, due to their remarkable binaural hearing and adaptive processing abilities. However, the ability of hearing-impaired listeners to understand human speech drastically plummets in the presence of reverberation and other interferences like noise or competing speakers [1]–[5]. Reverberation also severely affects the performance of sound localization algorithms [6] and speech recognition systems [7]–[9]. For these reasons, most communications systems now integrate speech *dereverberation* algorithms, aiming to remove the reverberant components picked up by the receiver and retrieve the anechoic (reverberation-free) speech. This increases the speech quality and intelligibility as well as transcription abilities in systems like, e.g. tele-conferencing, mobile phones and hearing devices [10]. Speech dereverberation can be considered a sub-task of speech *restoration* [11], which generally aims at restoring speech degraded by distortions like background noise, competing talk, bandwidth reduction, speech codec artifacts, etc.

Two dereverberation scenarios can be considered, namely the *informed* scenario, where the room impulse response (RIR) representing room acoustics is known; and the *blind* scenario where the RIR is unknown. Illustrations of a RIR, anechoic and reverberant signals are given in Figure 1.2. Although informed dereverberation is a much easier task in comparison to blind dereverberation, it is not trivial. This results from the difficulty of inverting non-minimum phase systems [12] and the sensitivity of L^p-optimization methods to RIR fluctuations caused by head movement, moving sources, etc. [P10], [13]–[15]. Because of the aforementioned sensitivity issues and the broader lack of knowledge of room acoustics in most scenarios, the blind case is of more practical relevance.

Traditional blind dereverberation methods (see Section 1.3) exploit statistical properties of speech, noise and reverberant signals [16]. When only one microphone is available, these algorithms mostly leverage temporal and spectral characteristics, leading to spectral enhancement [16], [17], cepstral processing [18] or linear prediction [19]–[21]. In the multi-channel case,

dereverberation methods can explicitly exploit spatial cues, resulting in beamforming [22], [23], convolutional beamforming [21], [24] or coherence weighting approaches [18], [25]–[29]. Probabilistic modelling plays an important role in speech dereverberation. Distributional assumptions, e.g. Gaussianity [21], sparsity [30]–[32] or low-rank non-negative matrix factorization (NMF) [33]–[35] can lead to maximum likelihood [21], [36], [37] and maximum a posteriori estimators [38], [39].

Although traditional approaches are robust to many scenarios, they struggle with non-stationary interferences, ubiquitous in real-life situations. Furthermore, many distributional assumptions considered in statistical methods often do not hold in practical use cases. These assumptions are indeed often meant to make the derivations of anechoic speech estimators tractable, rather than accurately describe the scenario. In the past decade, data-driven techniques using deep neural networks (DNNs) have gained prominence for audio processing tasks including speech enhancement and dereverberation [40]. DNNs are parametric estimators trained on large datasets to learn non-linear correlations and structures within speech data. At inference time, these learnt representations enable to discriminate between anechoic speech and reverberation, or to generate anechoic speech conditioned on reverberant measurements. Data-driven approaches therefore rely less on distributional assumptions than traditional methods but instead directly learn the signal properties and structures from data.

Predictive approaches using supervised learning objectives currently dominate the field of DNN-based speech restoration (see Section 1.4.1), given their impressive performance and conceptual simplicity [40], [41]. Using paired anechoic and reverberant speech data, these supervised predictive models learn to minimize a distance-like objective function between the processed reverberant speech and the anechoic reference. Such models can successfully perform dereverberation in the time-frequency domain [42]–[46] or directly on the waveform [47]–[49] and they can also handle multi-channel scenarios [50]–[52]. However, supervised predictive models suffer from poor generalization to unseen conditions such as new speakers or acoustic environments absent from their training conditions [P6], [P11].

For this reason, generative models and unsupervised learning are two directions of research being actively explored. Generative modelling aims at estimating and sampling from an intractable data distribution, such as the posterior of anechoic speech given a reverberant recording, rather than obtaining a single deterministic mapping between reverberant and anechoic speech. Generative models (see Section 1.4.2) boast interesting properties such as the ability of obtaining several valid estimates given a single reverberant condition, as well as a larger generalization ability to unseen conditions. In particular, diffusion-based generative models [53]–[55] encompass a class of powerful generative models that have been recently introduced for solving speech restoration tasks such as dereverberation. Diffusion models outperform earlier generative models and avoid many of their practical downsides, making them now serious contenders against predictive models for high-quality speech dereverberation [P7], [P8], [56].

Unsupervised learning is another paradigm in which models are trained to capture the structure of speech using only unlabeled data, in most cases anechoic speech. At test time, the resulting models are conditioned to align their estimate with the reverberant utterance. Unsupervised models benefit from a natural adaptability to various acoustic scenarios given the lack of supervision during training, among other advantages. However, their general performance is rarely on the level of supervised models in common scenarios, and training unsupervised models can be a challenging task.

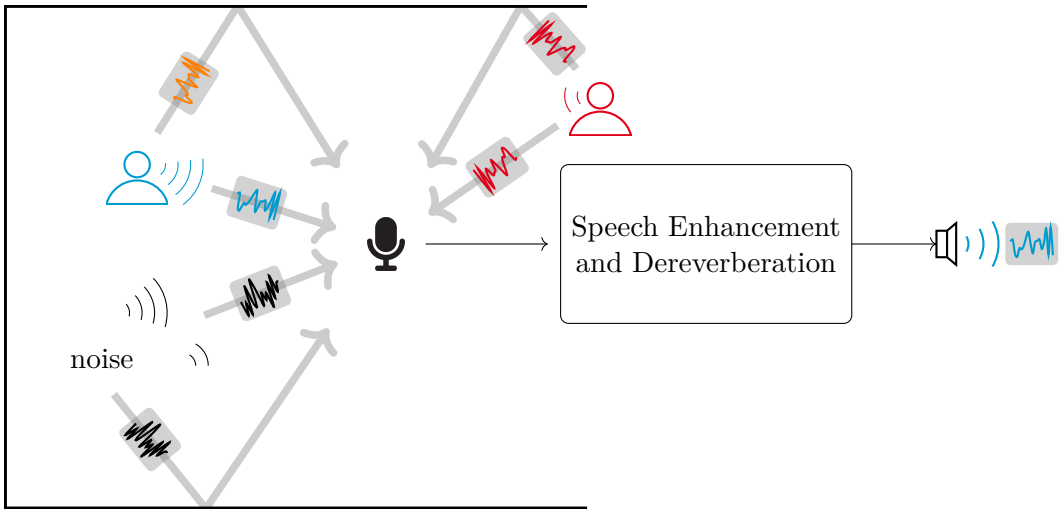


Figure 1.1: *Cocktail party scenario example, where a single microphone is available. One target speaker (blue) is present, along a noise source (black) and interfering speaker (red). All source signals are reverberated across the room and the system must extract the direct speech from the target source, removing the associated reverberant component (orange).*

It is important to note that data-driven and traditional speech dereverberation techniques have their respective drawbacks and strengths. A promising approach for achieving the best of both fields lies in finding hybrid methods, called *model-based* approaches (see Section 1.4.3). Model-based algorithms combine DNN-based estimators with prior knowledge of room acoustics, distributional assumptions or traditional signal processing methods. On the one hand, they are able to perform better than traditional algorithms thanks to the powerful non-linear regression and density estimation abilities of neural networks. On the other hand, leveraging algorithmic structure from successful traditional approaches and injecting prior knowledge help DNN-based techniques gain in robustness and efficiency.

In this thesis, we focus on two overlapping classes of DNN-based speech dereverberation techniques, namely conditional diffusion-based generative models and model-based algorithms. We first present model-based techniques using predictive models , then outline the advantages that conditional diffusion-based models offer for dereverberation. Both supervised and unsupervised learning paradigms are explored, and we demonstrate that diffusion-based techniques can benefit from the introduction of prior knowledge, therefore making a connection to previous model-based approaches.

In the remaining of this introduction, we first present some relevant background knowledge regarding room acoustics and human audition. An overview of traditional single- and multi-channel speech dereverberation methods is then given. We proceed to describe machine learning methods for dereverberation, with an emphasis on generative modelling and model-based methods. We conclude by describing historical and state-of-the-art methods for estimation of room acoustics as well as RIR simulation.

1.2 Room Acoustics and the Human Auditory Model

We provide here some background knowledge of acoustics, with a focus on signal models for reverberation in rooms. Some essential statistical and acoustic properties of reverberant

environments are presented. In the rest of this section, we broadly describe the human auditory system and give some insights into the characteristics of speech understanding and processing in the presence of hearing impairments.

1.2.1 Reverberant Signal Model

In static setups where the positions of a single sound source and an array with M microphones are fixed, reverberation is often modelled as a linear time-invariant (LTI) system. The reverberant utterance $y_m(n) \in \mathbb{R}$ recorded at the m-th microphone is obtained through a convolution of the anechoic, reverberant-free sound source $x(n) \in \mathbb{R}$ with the m-th microphone RIR $h_m(n) \in \mathbb{R}$ with length I:

$$y_m(n) = \sum_{i=0}^{I-1} h_m(i)x(n-i) = h_m(0)x(n) + r_m(n)\,, \tag{1.1}$$

where $r_m(n) \in \mathbb{R}$ denotes the corresponding pure reverberation component. Using a more compact notation, we can define the time domain convolution operator $*$ as:

$$y_m = h_m * x\,, \tag{1.2}$$

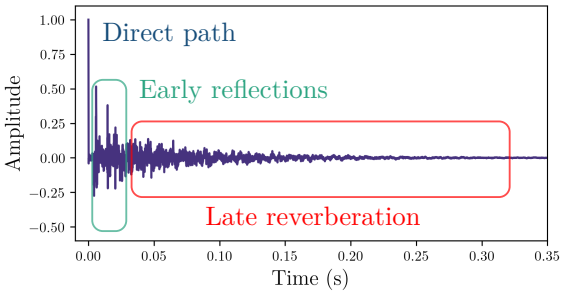
hereby ignoring the time variable.

The RIR can be separated into three components illustrated on Figure 1.2, namely the *direct path*, *early reflections* and *late reverberation*. The direct path is the first non-zero sample in the RIR, ignoring sampling artifacts and measurement noise. It corresponds to the path travelled by the direct sound wave between the sound source and the receiver. The early reflections denote the first few sound waves arriving after the direct sound, and form a region where times of arrival can be easily discriminated. Early reflections typically result in a frequency coloration and increase in the loudness of the perceived sound compared to the direct sound alone [59]. Furthermore, the auditory system can often leverage early reflections to extract spatial cues about the geometry of the room [59]. As time progresses, the number of reflections grows rapidly and the peaks in the RIR can no longer be attributed to separate arrivals of individual reflections, which is a phenomenon called *mixing*. Late reflections are the sound waves arriving to the microphone array once mixing is achieved, i.e. after the *mixing time*, which is usually taken to be between 24 and 50ms after the direct path depending on the definition and room dimensions [10], [60]. These late reflections result in a diffuse sound field whose energy density is uniform across the room, and are responsible for the degradation of speech quality and intelligibility [10].

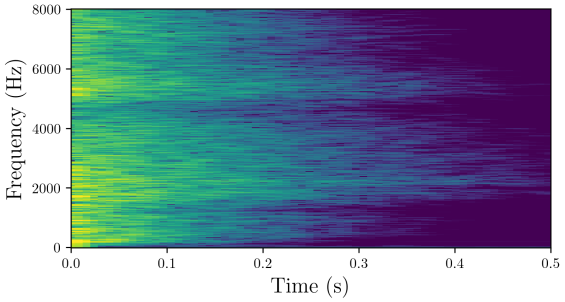
A well-known result of Fourier theory is that, when transposed to the Fourier domain, the convolutive model (1.1) can be expressed as a multiplication of the Fourier spectra of each signal. If signals are processed using the short-time Fourier transform (STFT) however, this statement no longer holds. Indeed, the window used for STFT analysis is of limited size, therefore the RIR is not guaranteed to fit in a single window. An accurate filtering model in the STFT domain is to consider the following set of multi-channel one-dimensional convolutions [61]:

$$Y_m(k,f) = \sum_{l=0}^{L-1} \sum_{\nu=-F_c}^{F_c} \tilde{H}_m(l,f,\nu)X(k-l,\nu)\,, \tag{1.3}$$

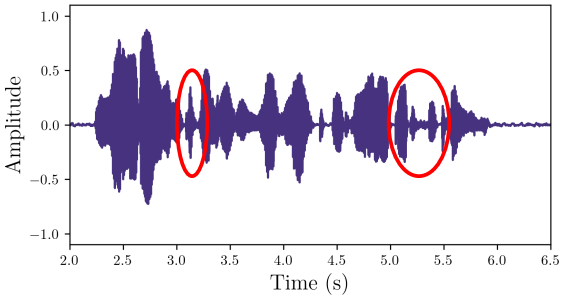
where f is the frequency bin, k is the time frame index, $Y_m(k,f) \in \mathbb{C}$ represents the STFT



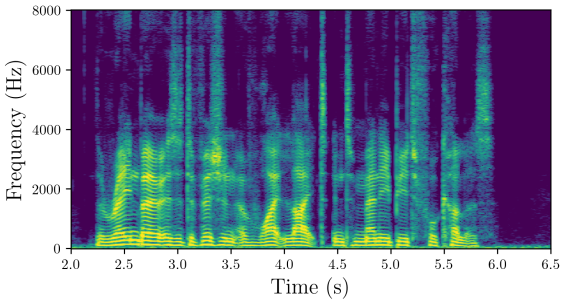
(a) RIR waveform



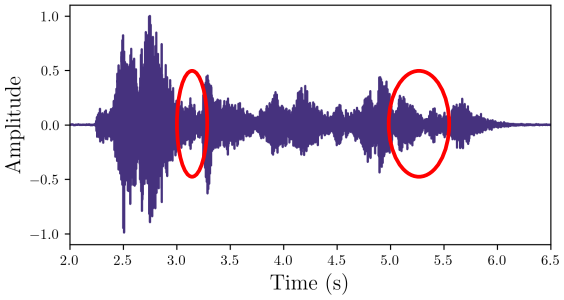
(b) RIR spectrogram



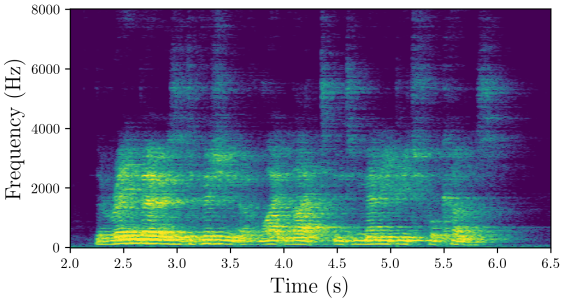
(c) Anechoic speech waveform



(d) Anechoic speech spectrogram



(e) Reverberant speech waveform



(f) Reverberant speech spectrogram

Figure 1.2: *Room impulse response (ARNI dataset [57]), anechoic speech signal (VCTK dataset [58]), and corresponding reverberant speech signal. Direct path, early reflections and late reverberation are highlighted on 1.2a. Some speech transients with sharp onsets and offsets are circled on 1.2c. These transients are completely smeared by reverberation in 1.2e.*

of $y_m(n)$ and $X(k,f) \in \mathbb{C}$ represents the STFT of $x(n)$. The filter $\tilde{H}_m(l,f,\nu) \in \mathbb{C}$ with time length L is interpreted as the response captured by the m-th microphone to a time-frequency impulse $\delta_{l,f-\nu}$ [61]. The sum over the index ν represents cross-band filtering up to F_c cross-bands, and the sum over the index l is a convolution along the time dimension in each frequency subband.

The *subband approximation* –or convolutive transfer function (CTF) model– ignores the effects of spectral leakage between neighbouring frequency bands, and is regularly used in dereverberation frameworks such as [39], [62], [63]. There, cross-band filters are discarded and a convolution is computed along the time dimension in each frequency band independently:

$$Y_m(k,f) = \sum_{l=0}^{L-1} H_m(l,f) X(k-l,f)\,, \tag{1.4}$$

where $H_m(k,f)$ is the STFT of $h_m(n)$.

The *narrowband approximation* implies going a step further and assuming that the length of the RIR filters $\mathbf{h}(n) = [h_1(n), \ldots, h_M(n)]^T$ is inferior to the STFT window length, thereby ignoring the STFT filter coefficients subsequent to the first one: $\forall l \geq 1\,,\, \mathbf{H}(l,f) \approx \mathbf{0}$. This yields the following filtering model, which is analogous to the result mentioned earlier when considering infinite windows for Fourier analysis:

$$Y_m(k,f) = H_m(0,f) X(k,f)\,. \tag{1.5}$$

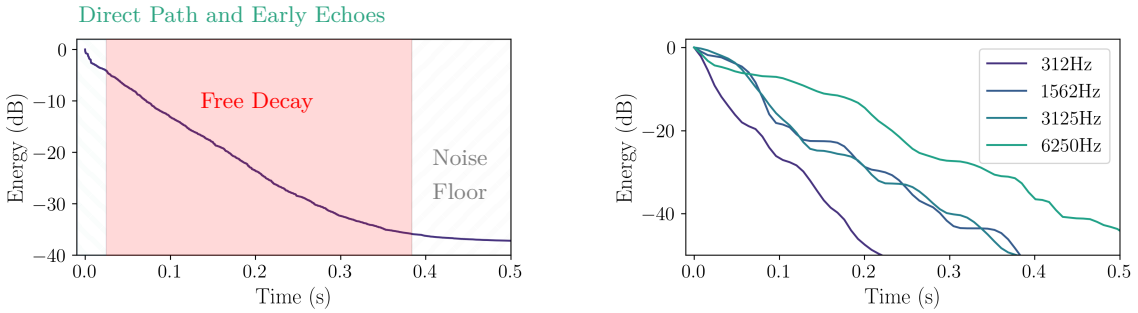
Anechoic or free-field scenarios are important cases where this narrowband approximation holds. In such cases, if the distance between the speaker and the microphone array is sufficiently large to neglect the inter-microphone level differences, the acoustic transfer function (ATF) $\mathbf{a}(f) := \mathcal{F}[\mathbf{h}](f) = \mathbf{H}(0,f) \in \mathbb{C}^M$ reduces to the following [64]:

$$\mathbf{a}(f) = a_1 \cdot \underbrace{\begin{bmatrix} 1 \\ e^{-2\pi j \Delta\tau_2 \frac{f_s \cdot f}{N_f}} \\ \cdots \\ e^{-2\pi j \Delta\tau_M \frac{f_s \cdot f}{N_f}} \end{bmatrix}}_{\mathbf{v}(f)}, \tag{1.6}$$

where $\mathcal{F}$ denotes the Fourier transform. The ATF $\mathbf{a}(f)$ factorizes into the complex gain a_1 of the reference microphone (here taken to be the first one) and the so-called steering vector $\mathbf{v}(f) \in \mathbb{C}^M$. The steering vector depends on the time difference of arrival (TDOA) $\Delta\tau_m \in \mathbb{R}$ between the reference and the m-th microphone, as well as the continuous frequency $\frac{f_s \cdot f}{N_f}$ corresponding to the discrete frequency bin f. The relation between discrete and continuous frequencies depends on the sampling frequency f_s and the STFT frame length N_f.

1.2.2 Statistical Properties of Reverberation and Acoustics

Given some assumptions on the considered medium (homogeneity, linearity, etc.), the propagation of sound through a material can be described using the second-order partial differential wave equation [59]. For reverberant environments, the sound field can be decomposed in a direct and a reverberant component. The energy of the direct component decreases following a inverse square law with respect to the distance [59]. In contrast, the reverberant sound



(a) Energy decay curve (EDC)

(b) Energy decay relief (EDR)

Figure 1.3: *Room impulse response decay characteristics.*
1.3a The RIR energy decays exponentially over the free-decay period, i.e. between 30 *and* 380*ms here until it hits the measurement noise floor.*
1.3b The rate of the exponential decay depends on the frequency because of frequency-dependent wall absorption profiles (see Figure 1.9).

energy is strictly independent of the distance between the sound source and the microphone but depends on the absorption of the material composing the room walls [59]. The lower the absorption of the materials, the higher the global reverberant field energy. Absorption profiles for various construction materials are shown on Figure 1.9.

An important acoustic indicator for studying and comparing rooms is the *reverberation time* T_{60}. It is defined as the time for the energy of a steady-state, diffuse sound field to decay by 60 dB [59]. The room must be excited with a broad-band signal until the resulting sound field reaches a steady state with diffuse properties, then the source is deactivated and the sound energy decay is measured. Measuring the reverberation time while the sound field is in a steady state , which guarantees the T_{60} is purely independent of the positions of the sound source and the microphone. The reverberation time can be approximated from the room characteristics using Sabine's formula [65]:

$$T_{60} \approx \frac{24 \ln(10)\, V}{c\, \alpha_{\text{Sabine}}\, A} \tag{1.7}$$

where ln is the natural logarithm, V is the volume of the room, $c = 344.0\,\text{m} \cdot \text{s}^{-1}$ is the speed of sound in air in normal pressure and temperature conditions, A is the total surface of room walls and α_{Sabine} is the average absorption coefficient of the room walls.

The direct-to-reverberant ratio (DRR) and C_{50} clarity index are also key metrics for characterizing room acoustics. The DRR is defined as the energy ratio between the direct path and the rest of the RIR:

$$\text{DRR}_m = 10 \log_{10} \frac{h_m^2(0)}{\sum_{n=1}^{I-1} h_m^2(n)}\,. \tag{1.8}$$

The C_{50} clarity index is given as the energy ratio between the 50 first milliseconds of the RIR, which comprise the direct path and most of the early reflections, and the remainder of the RIR:

$$C_{50,m} = 10 \log_{10} \frac{\sum_{n=0}^{\tau_{50}-1} h_m^2(n)}{\sum_{n=\tau_{50}}^{I-1} h_m^2(n)}\,, \tag{1.9}$$

with $\tau_{50} = 50$ms denoting a typical mixing time value separating early reflections from the late reverberation. As both the DRR and C_{50} take the direct path of the RIR into consideration, these metrics explicitly depend on the relative positioning of the source and the microphone, and are not just room-dependent like the T_{60} reverberation time.

The evolution of the RIR energy over time can be assessed computing the following Schroeder integral (here discretized) called energy decay curve (EDC):

$$\mathrm{EDC}_m(n) = \sum_{i=n}^{I-1} h_m^2(i)\,. \tag{1.10}$$

It can be observed on Figure 1.3a that the EDC corresponding to the RIR illustrated on Figure 1.2a decays exponentially during the so-called *free decay period* over which the sound field in the room is diffuse. In the presented case, this free decay period approximately runs from 35ms (after the direct path and early reflections) until 380ms (where the energy then decays beyond the noise floor of the measurement, here about -36dB below the direct path). Coarse estimation of T_{60} reverberation times can also be carried out directly on the EDC plot [10]. This characteristic has been leveraged by Polack to propose the following stochastic model for the diffuse part of a room impulse response [60]:

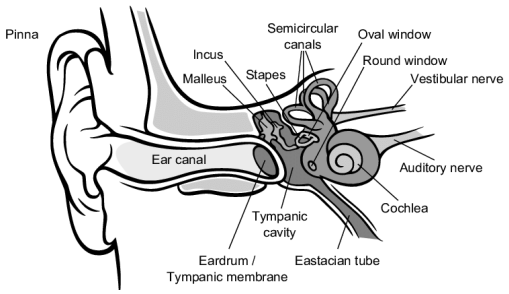
$$h(n) \approx b(n) \exp\left(-\frac{3\ln(10)}{T_{60}}\, n\right)\,, \tag{1.11}$$

where $b(n)$ is a zero-mean stationary Gaussian process, i.e. an infinite collection of zero-mean Gaussian random variables, whose covariance matrix $\mathbb{E}[b(n)b(n^{'})]$ depends only on the time lag $n - n^{'}$. This stochastic model is valid for frequencies higher than the *Schroeder frequency*, below which the room acoustics must be analyzed through harmonic solutions to the wave equation, because of the prominence of resonant modes [59]. A simple expression of the Schroeder frequency can be obtained via the T_{60} reverberation time and room geometry:

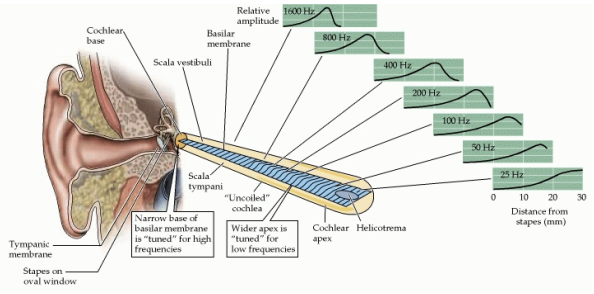
$$f_g = C\sqrt{\frac{T_{60}}{V}}\,, \tag{1.12}$$

where V is the volume of the room and $C \approx 2000\,(\mathrm{m}\cdot\mathrm{s}^{-1})^{3/2}$. For instance, the Schroeder frequency of an office room with dimensions $8 \times 6 \times 3$ meters and a reverberation time of $T_{60} = 0.4$s is of approximately 105Hz. As mentioned earlier, the validity of Polack's statistical model (1.11) is also temporally limited to the free-decay period, i.e. when mixing is achieved.

Another key observation becomes evident when computing the energy decay relief (EDR), which is obtained by isolating some frequency bands in the RIR and computing the resulting EDC in each band [66]. As visualized on Figure 1.3b, we observe that the decay of the sound energy depends on the frequency band, which naturally results from the frequency-dependent behaviour of the room walls materials. The general trend is that materials absorb high frequencies more efficiently than low frequencies, as shown on Figure 1.9. Therefore, the reverberation time at low frequencies is most of the time longer than for higher frequency bands. This motivates the time-frequency model of late reverberation proposed in [P12].



(a) Overview of human auditory model. Taken from [68]. Original graphics by C. L. Brockmann.



(b) Uncoiled cochlear and basilar membrane. Taken from [69].

Figure 1.4: *Schematics of human auditory model*

1.2.3 Human Auditory System and Signal Processing for Hearing Devices

The human auditory system is an electromechanical transducer that transforms the sound pressure signal at the pinna to an electrical signal in the auditory brainstem [67]. The acoustic wave first travels air down the auditory canal until the eardrum (see Figure 1.4a). The middle ear ossicles (malleus, incus, stapes) then transmit the sound until the cochlea, a snail-shaped organ in the inner ear composed of various chambers. Vibrations of the chamber inner fluid translate to a longitudinal excitation of the basilar membrane, where different hair cells will be stimulated depending on the frequencies composing the audio signal (see Figure 1.4b). Hair cells transform the nature of the signal from a longitudinal mechanical wave into an electrical stimulus through modifications of the local polarization and opening of ionic channels. The sound information is then transmitted via the auditory nerve all the way to the auditory brainstem through multiple elaborate pathways, along which diverse cues such as location, intensity or pitch are extracted and processed by the cortical areas [67].

Hearing loss refers to any impairment affecting one or more components of the auditory system described above. The World Health Organization (WHO) estimates that in 2021 nearly 20% of the global population lived with a hearing loss, with 430 million individuals (over 5%) requiring a proper rehabilitation [70]. These figures are on the rise, as the WHO projects that as many as 700 million people (i.e. 1 individual out of 10) could suffer from disabling hearing loss by 2050. Hearing losses are naturally depending on the age group: nearly 30% of people aged 60 years or above have at least some level of hearing loss [70]. This partially explains the impressive growth of the fraction of hearing-impaired listeners among humans, as the global population is rapidly ageing. One of the typical age-related degradations is the reduction of hair cell functions, leading most of the time to a loss of high-frequencies in the perceived sounds [71]. If such hearing impairment can be considered mild, severe hearing loss can be caused by more traumatic physiological and toxological events, leading to the destruction of middle ear bones, heavy deficits in inner hair cells or even damages in the auditory nerve synapse [67]. All degrees of hearing impairment translate to a loss of temporal and spectral resolution of the auditory pathway. This results in reduced speech intelligibility and increased listening effort, especially in noisy and reverberant situations [1], [3]–[5]. Concretely, Dillon [2] predicts a loss of about 4 to 10 dB for speech-in-noise reception thresholds of hearing-impaired listeners. This form of perceptual hearing loss can not be solved by sole amplification of the incoming signal, but requires an effective removal of the speech interferences. Interestingly, Beck et

al. [72] reported in 2018 that 26 million Americans had hearing difficulty and/or difficulty understanding speech-in-noise, despite having clinically normal thresholds in clean conditions. This suggests that general diagnosis of hearing impairments might be more difficult than anticipated, and that tools like, e.g. pure-tone audiograms or clean-only speech understanding tests could be unfit to detect such cases.

Surgical treatment of hearing losses is delicate or even impossible in most cases, and solutions instead rely on embedded devices such as behind-the-ear hearing aids and cochlear implants, depending on the cause and severity of the hearing impairment. A hearing aid is an external device that processes the sound incoming at the integrated microphones, and eventually transmits the processed sound to the auditory canal. Thus, it can only treat partial hearing losses where the inner ear is not significantly damaged. When the cochlea is severely degraded or even destroyed, a cochlear implant is required. Cochlear implants process the captured acoustic wave in the external part of the device situated outside of the skull. Then, an implant inserted into the temporal bone directly excites the auditory nerve fibers using a coding scheme mimicking the inner ear's behaviour.

Audio processing in such hearing devices was historically limited to operations like multi-band compression or equalizing, since computational capacities of hearing devices are limited by their low battery life. More recent devices allowed the integration of traditional speech enhancement and dereverberation algorithms (see Section 1.3 for more details), e.g. multi-microphone beamforming [3], [73], [74], coherence weighting [25], [26], statistical spectral enhancement [17], [75], [76] or spectral subtraction [77], [78]. Evaluating the impact of speech enhancement and dereverberation algorithms on hearing-aided and cochlear-implanted listeners is an industrious yet paramount task. It has been shown that single-channel noise reduction traditional algorithms improve the intelligibility of speech in noise for normal-hearing and hearing-aided listeners [79]–[81] as well as for cochlear-implanted users [82]–[84]. Interestingly, a listening study performed on cochlear-implanted listeners [85] suggests that they are less averse to speech distortions than hearing-aided users but are more sensitive to residual interferences in the speech signal. For reverberation in particular, it was shown that early reflections might benefit hearing-aided listeners [86] (like normal-hearing people) but not necessarily patients equipped with cochlear implants [87].

As new developments in hardware research have lead to increased computational power and battery life, more ambitious audio processing are now considered, for instance relying on DNNs [88], [89]. However, DNN-based speech processing is computationally demanding when compared to traditional algorithms. Therefore, special care must be given to keeping processing within hardware limits. For hearing device applications, the global latency should be kept inferior to a threshold determined by listener-dependent requirements with respect to audiovisual cues and self speech hearing. Multiples studies demonstrate the importance of visual cues for speech understanding in noisy situations. Sumby and Polack show that in extremely noisy scenarios (down to -30dB signal-to-noise ratio (SNR)), the word intelligibility score can improve by as high as 40% when switching from pure audio to audiovisual cues [90], [91]. Normal-hearing listeners and hearing device users can tolerate levels of audiovisual asynchronicity up to $\sim$ 200ms (worsening with age and hearing impairment severity [92]), which theoretically allows for large latencies. However, the primary limiting factor occurs when listeners using hearing aids (as opposed to cochlear implant users) hear their own voices. Hearing-aided listeners, unlike cochlear-implanted users, have residual hearing, which implies that they hear their own voice through multiple pathways. The first pathway for self speech hearing results from air leakage around the hearing aid earmold. The second is

natural bone conduction of the sound wave, and the last corresponds to the signal travelling via the air from the mouth to the hearing aid microphone. These three signals add up with different delays at the input of the auditory system, and therefore the auditory system can only seamlessly integrate these signals if the delay caused by hearing aid processing (and speech production itself) is low enough. Stone and Moore show that this drastically limits the allowed latency in hearing aids to $\sim$ 20ms [93]. Hearing aid manufacturers consider even lower latency targets. For instance, the rules of the Clarity challenge for machine learning-based speech noise reduction in hearing aids only allow algorithms with a global latency below 5ms [94].

The global latency is the sum of the *algorithmic* and *hardware* latencies. The algorithmic latency defines how much time is needed to output an audio frame, starting from the moment a corresponding input frame enters the system, given an infinite computing power. In other terms, it relates to how much time is needed to fill the output buffer of the system. Limiting the algorithmic latency amounts to only using a limited amount of future information for processing, thereby stating that processing should be *causal*. In algorithms using representations provided by, e.g. STFT or Gammatone filterbanks [95], the algorithmic latency is dictated by the length of the synthesis window, i.e. the length of the pseudo-inverse filters processing the internal representation back to the original waveform domain. Using asymmetric analysis/synthesis windows in filterbanks has been investigated as an elegant solution to reduce algorithmic latency without sacrificing spectral resolution [96], which is paramount for speech enhancement and dereverberation. In DNN-based processing, guaranteeing causal processing involves for instance asymmetric padding in convolutional layers, adaptive normalization strategies [97], or adopting recurrent network architectures.

The hardware latency relates to the computational complexity of the algorithm on the considered hardware. We define here the real-time factor (RTF) as the ratio between the time used for processing some audio with respect to the duration of this input signal:

$$\text{RTF} = \frac{\text{Processing Time}}{\text{Input Signal Duration}}\,. \tag{1.13}$$

The RTF should be kept below one for the algorithm not to introduce any hardware latency. If the algorithmic latency can be allowed to be non-zero (but sufficiently small), this is not the case for the hardware latency, as any remaining time dedicated to computations above the specified target will accumulate over time, thereby inevitably exceeding the allowed global latency at some point. This is arguably the most critical point on which DNN-based algorithms behave poorly. However, many techniques are being developed for solving that issue, such as the design and optimization of DNN-oriented hardware, or the reduction of neural network size through model pruning and quantization [98] or knowledge distillation [99]. One should further privilege efficient architectures requiring few operations per second, for instance preferring recurrent structures with linear computational complexity with respect to the input sequence length [100]–[102], compared to attention mechanisms having quadratic complexity [103]. Developing so-called *model-based* algorithms, i.e. hybrid methods integrating traditional signal processing with elements of deep learning can also be an interesting lead for curbing computational complexity (see Section 1.4.3).

Finally, one last constraint is related to memory. In the context of DNNs, this means that the memory occupied by all internal buffers as well as the model weights should be be kept within the capacity of the device. Engineering techniques for shrinking the memory footprint of

neural networks include pruning and quantization [98] and knowledge distillation [99].

1.3 Traditional Speech Dereverberation

We review in this section so-called traditional speech dereverberation algorithms, that do not rely on data for estimating the anechoic speech. Traditional methods exploit temporal, spectral and spatial cues of anechoic and reverberant speech. They can also leverage statistic models of reverberation and distibutional assumptions of speech signals, as well as signal model approximations.

1.3.1 Informed Inverse Filtering

Traditional informed dereverberation methods have access to the complete knowledge of the RIR. They are usually based on either exact inverse filtering or approximate filter optimization.

Exact inverse filtering methods generally operate in the Fourier domain, relying on the fact that the convolutive signal model (1.1) is perfectly represented by the narrowband assumption (1.5) if the STFT analysis window is longer than the RIR. An exact inverse filter can therefore be taken as the inverse of the Fourier spectrum of the RIR, transformed back to the original time-domain. However, most real RIR signals are mixed-phase, which means the inverse filter computed following the steps described above cannot be both causal and stable [12]. For that reason, naive inverse filtering in the Fourier domain leads to problematic artifacts such as pre-echoes and feedback-like tones. Kodrasi et al. [104] propose to regularize the inverse filter by forcing the zeros of the RIR z-transform into the unit circle. However, this approach is intrinsically limited as the resulting inverse filter is only an approximation of the real inverse system. Consequently, a reasonably small regularization factor is chosen, and residual artifacts are attenuated using a traditional speech enhancement scheme. In [105], a method based on homomorphic signal analysis is proposed, directly trying to address the problems caused by the mixed-phase nature of RIRs. There, the mixed-phased system is decomposed into a minimum-phase / maximum-phase or minimum-phase / all-pass filter pair, and the inverse filter correspondingly derived. However, compared to a simple least-squares approximation of the inverse filter (see thereafter), the proposed approach lacks accuracy [105]. Stability issues can also be alleviated by computing an exact filter using recordings from multiple microphones following [106]. However, this implies having access to such multi-channel recordings, and a further requirement is that RIRs should have no common zero across channels in the z-plane [106].

Other informed methods instead rely on approximate inverse filtering, which is better understood under the perspective of solving an *inverse problem*. The anechoic speech is obtained by optimizing the following objective function:

$$\hat{x} = \arg\min_{x \in \mathbb{R}} \mathcal{J}(h * x, y) + \mathcal{R}(x) \tag{1.14}$$

where the time index n is omitted for the sake of readability. The functional $\mathcal{J}$ measures the discrepancies between the reverberant measurement y and its estimated reconstruction under the signal model (1.2). The regularizer $\mathcal{R}$ imposes some prior knowledge on the search quantity, i.e. anechoic speech, in order to shrink the size of the space of solutions. This is necessary because dereverberation is a form of *deconvolution*, which is a notoriously *ill-posed* inverse problem. This means that several anechoic speech solutions are $\mathcal{J}$-optimal given

a single reverberant recording. This characteristic makes most inverse problem solvers fail to converge to a suitable solution, if no proper prior is imposed on the anechoic speech. Traditional approaches only dispose of unstructured priors that provide basic information about the anechoic speech characteristics. In the image deconvolution literature, traditional priors include the total variation norm [107], [108], the L^1 norm for sparsity promotion [109], or the classical L^2 norm [110].

Early dereverberation works consider minimizing a well-behaved least-squares reconstruction objective [105], [111], [112], while other L^p-based optimization rules are presented in [113], [114]. Unfortunately, such methods (as well as exact inverse filtering) are quite sensitive with respect to observation noise and fluctuations in the RIRs [P11], [13]–[15]. Similar to what is proposed in [104] for direct computation of a regularized inverse filter, Hikichi et al. [15] derive a regularized objective for least-squares based approximate inverse filtering, penalizing the filter energy under various frequency profiles of perturbation noise.

Given the sensitivity issues mentioned above and, more importantly, the general lack of knowledge of room acoustics in most real-life scenarios, we will focus on blind dereverberation approaches in the remainder of this section.

1.3.2 Blind Single-Channel Dereverberation

In this subsection we omit the microphone index m and the corresponding bold face notation as only one microphone is considered available. Early blind single-channel dereverberation techniques rely on auto-regressive linear prediction for signal analysis:

$$\forall n \in \mathcal{V} \,:\, y(n) = \sum_{i=0}^{P-1} c_{\mathcal{V}}(i)\, y(n-i) + e(n)\,, \tag{1.15}$$

where $\{c_{\mathcal{V}}(i)\}_{l=0}^{P-1}$ are the P-th order linear prediction coefficients for the currently analyzed time segment $\mathcal{V}$ and $e(n)$ is the linear prediction residual. In [19] for instance, the linear prediction residual is weighted in order to emphasize regions with high signal-to-reverberation ratio. Gillespie et al. [20] propose instead to maximimize the residual's kurtosis, i.e. its centered third-order moment, based on the observation that anechoic signals have a higher kurtosis than reverberant signals.

Another strategy for blind dereverberation is to directly exploit the spectral information. These *spectral enhancement* techniques derive a spectral magnitude gain function $G(k,f) \in \mathbb{R}^+$, designed as the ratio of the anechoic speech and reverberant speech magnitudes, to yield the processed output magnitude:

$$|\hat{X}(k,f)| = G(k,f) \cdot |Y(k,f)|\,. \tag{1.16}$$

Such methods include for instance spectral subtraction [77], [115] and statistical spectral enhancement [16], [17]. Spectral subtraction methods aim to estimate the reverberation signal short-term power spectral density (PSD) $\gamma(k,f) := \mathbb{E}[|R(k,f)|^2]$ where $R = Y - X$ is the pure reverberation component. The estimated PSD square root is then subsequently subtracted from the reverberant mixture magnitude, leading to the following spectral gain formula:

$$G(k,f) = 1 - \frac{\sqrt{\hat{\gamma}(k,f)}}{|Y(k,f)|} \tag{1.17}$$

The reverberant PSD can be estimated through diverse tools such as, e.g. Polack's statistical model (1.11) in [77], or linear prediction in [115]. A known weakness of spectral subtraction is that the dereverberated output often contains random tonal artifacts. These artifacts result from the fact that some reverberation components in the mixture magnitude spectrum $|Y|$ can go below the average estimate given by $\hat{\gamma}$. Since the gain function must remain positive *as per* its definition in (1.16), the naive solution is to rectify its values: $\tilde{G}(k,f) = \max(0, G(k,f))$. However, this non-linear rectification causes unpleasant artifacts, which are referred to as *musical noise*. Some techniques have been developed to alleviate this issue: in [77] for example, the PSD estimate $\hat{\gamma}$ is time-averaged, or a spectral floor is used to limit the values of the gain function G. Both measures tend to limit the dynamics of the applied spectral gain, therefore mitigating musical noise but inevitably leaving some residual reverberation in the processed signal.

Some early works also explore subband envelope filtering [116], [117] based on the CTF assumption (1.4) to increase the modulation depth of the resulting signal.

In contrast, statistical spectral enhancement methods [16], [17], [118] derive the spectral gain function G by minimizing a distortion measure between anechoic and processed signals. The involved derivations rely on distributional assumptions of the speech and reverberation signals. For instance in [75], the speech, noise and reverberation are modelled using Gaussian distributions, the distortion measure is the squared error and a log-spectral amplitude estimator is used, leading to the minimum mean square error log-spectral amplitude estimator (MMSE-LSA) under speech probability uncertainty [119]. Another celebrated example is the Wiener filter, i.e. the least-squares estimator of target speech when considering anechoic speech and reverberation as uncorrelated Gaussian signals [120]. Assuming $X(k,f) \sim \mathcal{N}_{\mathbb{C}}(0; \sigma_x^2(k,f))$ and $R(k,f) \sim \mathcal{N}_{\mathbb{C}}(0; \sigma_r^2(k,f))$ independently for each time and frequency bin, the Wiener filter is obtained as:

$$G_{\text{Wiener}}(k,f) = \frac{\sigma_x^2(k,f)}{\sigma_x^2(k,f) + \sigma_r^2(k,f)}\,, \tag{1.18}$$

under the uncorrelation constraint $\mathbb{E}\left[X(k,f)R^*(k,f)\right] = 0$.

Other signal domains can also be exploited for dereverberation. For instance in [18], [121] spectral gains are smoothed using real-valued speech cepstra, thereby reducing musical noise without smearing speech onsets and offsets. Some works also leverage the wavelet representation of reverberant speech, for either clustering of linear prediction coefficients [122] or wavelet Wiener filtering [123].

1.3.3 Blind Multi-Channel Dereverberation

When multiple microphones are available, spatial information can be exploited. Given the geometrical nature of the reverberation phenomenon, spatial cues constitute precious information for the retrieval of anechoic speech. For instance, single-channel linear prediction methods can easily extend to multi-channel settings by, e.g. averaging the microphone signals before applying the linear prediction analysis [124], [125].

Linear Filtering

A wide fraction of multi-channel dereverberation methods perform beamforming in the STFT domain [22], [23], [28], [126]. Beamformers are linear spatial filters, directly applied to the multi-channel reverberant mixture short-time spectrum $\mathbf{Y}(k,f) = [Y_1(k,f), \ldots, Y_M(k,f)]$. Here we consider a multi-input single-output setting, where the goal of the linear spatial

filter is to retrieve the anechoic speech at the reference microphone, here taken to be the first one:

$$\hat{X}_1(k,f) = \mathbf{W}^H(k,f)\mathbf{Y}(k,f)\,. \tag{1.19}$$

Deriving the beamformer weights $\mathbf{W} \in \mathbb{C}^M$ often requires some knowledge of the speech and interfering signals covariance matrices $\Phi_X(k,f) := \mathbb{E}[\mathbf{X}(k,f)\mathbf{X}^H(k,f)] \in \mathbb{C}^{M\times M}$ and $\Phi_R(k,f) := \mathbb{E}[\mathbf{R}(k,f)\mathbf{R}^H(k,f)] \in \mathbb{C}^{M\times M}$, respectively. A seminal example is the minimum variance distortionless response (MVDR) beamformer [120], which according to its name minimizes the variance of the residual reverberation at the filter output while introducing no distortion to the target signal:

$$\mathbf{W}_{\text{MVDR}}(k,f) = \arg\min_{\mathbf{W}\in\mathbb{C}^M} \mathbf{W}^H(k,f)\Phi_R(k,f)\mathbf{W}(k,f) \quad \text{s.t.} \quad \mathbf{W}^H(k,f)\mathbf{v}(f) = 1\,, \tag{1.20}$$

where $\mathbf{v}(f)$ is the steering vector with respect to the first microphone, defined in (1.6). Optimizing this objective with the Lagrange multiplier method yields the following expression for the MVDR beamformer weights [120]:

$$\mathbf{W}_{\text{MVDR}}(k,f) = \frac{\Phi_R^{-1}(k,f)\mathbf{v}(f)}{\mathbf{v}^H(f)\Phi_R^{-1}(k,f)\mathbf{v}(f)}\,. \tag{1.21}$$

Dietzen et al. [23] integrates another classical beamformer called generalized sidelobe canceller (GSC) together with an adaptive filtering technique for speech dereverberation separation and noise reduction. Kuklasinski et al. [22] run a comparison of various PSD estimation strategies for multi-channel Wiener filtering, which is the concatenation of the MVDR beamformer with a single-channel Wiener post-filter [127]. Various statistical multi-channel techniques integrating beamformers are proposed and compared for joint denoising and dereverberation in [128], including the multi-channel Wiener filter and a regularized partial channel equalization for dereverberation [129]. In [126], an MVDR beamformer is combined with a sophisticated single-channel post-filter involving spectral and cepstral processing [130]–[132].

Other post-filtering strategies can be derived by aiming to increase the cross-channel coherence of the dereverberated speech [18], [25]–[29]. The cross-channel coherence function is defined as the pairwise normalized cross-power spectrum between signals at different channels [73]:

$$\Gamma^x_{m,n}(f) = \frac{\mathbb{E}[X_m(k,f)X_n^*(k,f)]}{\sqrt{\mathbb{E}[|X_m(k,f)|^2]\mathbb{E}[|X_n(k,f)|^2]}} \tag{1.22}$$

where the expectation is taken on the time dimension. A relevant example for dereverberation is the coherence matrix of an isotropic diffuse sound field, which is often chosen to represent the pure reverberation component in isotropic acoustic environments [27]:

$$\Gamma^{\text{diff.}}_{m,n}(f) = \text{sinc}\left(\frac{2\pi f d_{m,n}}{c}\right)\,, \tag{1.23}$$

where $d_{m,n}$ is the distance between the m-th and n-th microphone. As they travel across the room and get reflected by walls (occurring unknown phase jumps), reverberated signals lose their coherence [59]. On the contrary, direct speech signals retain a high coherence across the microphone array. Therefore, maximizing coherence in the output speech can be seen as a proxy for reducing early echoes and residual reverberation. Allen et al. [18], [25] directly compute a Wiener-like filter using coherence matrix estimates, whereas [26]–[28] harness model

assumptions for the reverberant sound field. While [26] models the reverberation coherence as a fully incoherent sound field (i.e. microphone-independent sensor noise), [27] makes the more suitable assumption of an isotropic diffuse sound field (1.23). Jeub et al. [28] further refine the diffuse model from [27] to include a binaural model of the human head, thereby accounting for head-shadowing effects. Schwartz et al. [37] compare different estimators for the DRR (1.8) based on coherence matrices, using isotropic diffuse and two-dimensional-isotropic noise field models. They subsequently derive a spectral subtraction algorithm based on the obtained DRR estimate. The diffuse sound field coherence matrix in (1.23) can also be used for multi-channel beamforming. When used in place of the interference covariance matrix in the MVDR beamformer (1.21), it results in the so-called super-directive beamformer [133]:

$$\mathbf{W}_{\mathrm{SDB}}(f) = \frac{\left(\mathbf{\Gamma}^{\mathrm{diff.}}(f)\right)^{-1}\mathbf{v}(f)}{\mathbf{v}^{H}(f)\left(\mathbf{\Gamma}^{\mathrm{diff.}}(f)\right)^{-1}\mathbf{v}(f)}\,. \tag{1.24}$$

Statistical Methods and the WPE Algorithm

As of the late 2010s, the field of multi-channel dereverberation has been increasingly dominated by statistical estimation techniques relying on maximum likelihood (ML) objectives. ML optimization consists in maximizing the likelihood of the reverberant measurement $\mathbf{Y}$ under a parametric model given by the probability density function (PDF) $p(\mathbf{Y}|\theta)$ where $\theta \in \Theta$ are the parameters of the distribution. Here the objective is equivalently formulated as a minimization of the negative log-likelihood function:

$$\theta^{*} = \arg\min_{\theta\in\Theta} \mathrm{NLL}(\theta) = \arg\min_{\theta\in\Theta} \; -\log\, p(\mathbf{Y}|\theta)\,, \tag{1.25}$$

where we omit the time and frequency indexes for simplicity. The ML objective (1.25) often does not have a closed-form solution, either because of the composition of the parameter set θ or the distributional form of $p(\cdot|\theta)$. Therefore some iterative algorithms like the expectation-maximization (EM) algorithm [134] or coordinated gradient descent [21] are used.

A landmark ML-based algorithm for multi-channel dereverberation is the celebrated weighted prediction error (WPE) algorithm [21]. WPE combines a subband auto-regressive reverberant signal model with a ML objective under a time-varying Gaussian prior of the target speech component. Consider the CTF signal model following (1.4):

$$Y_m(k,f) = \sum_{l=0}^{L-1} H_m(l,f)X(k-l,f) \tag{1.26}$$

$$= \underbrace{\sum_{l=1}^{\Delta-1} H_m(l,f)X(k-l,f)}_{D_m(k,f)} + \underbrace{\sum_{l=\Delta}^{L-1} H_m(l,f)X(k-l,f)}_{\tilde{R}_m(k,f)}\,, \tag{1.27}$$

where L is the length of the RIR short-time spectrum $H_m(k,f) \in \mathbb{C}$. Here $D_m(k,f) \in \mathbb{C}$ denotes the short-time spectrum of the desirable component, which contains the direct path and some early reflections, depending on the value of the prediction delay $\Delta \in \mathbb{N}^{*}$. The short-time spectrum $\tilde{R}_m(k,f) \in \mathbb{C}$ corresponds to the reverberant component we wish to remove. Nakatani et al. [21] consider the following linear prediction analysis, assuming that the reverberant component can be well represented by a P-th order multi-channel auto-regressive

model with unknown frequency-dependent coefficients $\{\mathcal{G}(l,f) \in \mathbb{C}^{M\times M}\}_{l=0}^{P-1}$:

$$Y_m(k,f) = D_m(k,f) + \sum_{\tilde{m}=1}^{M}\sum_{l=0}^{P-1} \mathcal{G}^*_{m,\tilde{m}}(l,f) Y_{\tilde{m}}(k-\Delta-l,f)\,. \tag{1.28}$$

Using a matrix notation, the auto-regressive model (1.28) can be rewritten as:

$$\mathbf{Y}(k,f) = \mathbf{D}(k,f) + \tilde{\mathcal{G}}^H(f)\mathcal{Y}_\Delta(k,f)\,, \tag{1.29}$$

where $\mathcal{Y}_\Delta(k,f)$ denotes the delayed multi-frame reverberant signal window:

$$\mathcal{Y}_\Delta(k,f) = [\mathbf{Y}(k-\Delta,f),\, \mathbf{Y}(k-\Delta-1,f),\, \ldots,\, \mathbf{Y}(k-\Delta-P+1,f)]^T \in \mathbb{C}^{MP}\,, \tag{1.30}$$

and $\tilde{\mathcal{G}}(f) \in \mathbb{C}^{MP\times M}$ is obtained by concatenating the WPE multi-frame coefficients $\mathcal{G}(f)$ on the first axis. The desired component $\mathbf{D}(k,f)$ is assumed to follow a zero-mean time-varying Gaussian prior with microphone-independent variance $\sigma_d^2(k,f)$:

$$\mathbf{D}(k,f) \sim \mathcal{N}_\mathbb{C}(\mathbf{0}, \sigma_d^2(k,f)\mathcal{I})\,, \tag{1.31}$$

where $\mathcal{I} \in \mathbb{R}^{M\times M}$ is the identity matrix. The auto-regressive model (1.29) and Gaussian assumption (1.31) lead to the following negative log-likelihood function with parameters $\theta = \{\tilde{\mathcal{G}}, \sigma_d^2\}$:

$$\mathrm{NLL}(\tilde{\mathcal{G}}, \sigma_d^2) = \mathbb{E}\left[-\log \mathcal{N}_\mathbb{C}\left(\tilde{\mathcal{G}}^H(f)\mathcal{Y}_\Delta(k,f)\,,\, \sigma_d^2(k,f)\mathcal{I}\right)\right] \tag{1.32}$$

$$= \mathbb{E}\left[\frac{1}{2\sigma_d^2(k,f)}||\mathbf{Y}(k,f) - \tilde{\mathcal{G}}^H(f)\mathcal{Y}_\Delta(k,f)||_2^2 + \frac{1}{2}\log\sigma_d^2(k,f)\right] + \text{const.} \tag{1.33}$$

The joint ML optimization objective (1.32) does not have a closed-form solution, and therefore a coordinated gradient descent approach is used to estimate the parameters. This consists in iterating between updates for the auto-regressive coefficients $\tilde{\mathcal{G}}$ and the desired component PSD σ_d^2. The variance update computes the variance σ_d^2 assuming the filter coefficients are fixed, using a simple periodogram estimate:

$$\sigma_d^2(k,f) = ||\mathbf{Y}(k,f) - \tilde{\mathcal{G}}^H(f)\mathcal{Y}_\Delta(k,f)||_2^2\,. \tag{1.34}$$

Subsequently, the filter update minimizes (1.32) with respect to the WPE filter coefficients $\tilde{\mathcal{G}}$, considering the speech variance known. A least-square derivation step yields the following expression for the WPE filter:

$$\tilde{\mathcal{G}}(f) = \mathcal{P}^{-1}(f)\mathcal{Q}(f)\,, \tag{1.35}$$

where $\mathcal{P}(f) \in \mathbb{C}^{MP\times MP}$ and $\mathcal{Q}(f) \in \mathbb{C}^{MP\times P}$ are respectively the time-averaged variance-normalized correlation matrix and vector of the delayed signal window:

$$\mathcal{P}(f) = \mathbb{E}\left[\frac{\mathcal{Y}_\Delta(k,f)\mathcal{Y}_\Delta^H(k,f)}{\sigma_d^2(k,f)}\right]\,, \tag{1.36}$$

$$\mathcal{Q}(f) = \mathbb{E}\left[\frac{\mathcal{Y}_\Delta(k,f)\mathbf{Y}^H(k,f)}{\sigma_d^2(k,f)}\right]\,. \tag{1.37}$$

The dereverberated speech estimate is then obtained after these two updates following

(1.29):

$$\hat{\mathbf{D}}(k,f) = \mathbf{Y}(k,f) - \tilde{\mathcal{G}}^H(f)\mathcal{Y}_\Delta(k,f)\,. \tag{1.38}$$

Numerous extensions of the WPE method have been proposed in the following years [24], [30]–[33], [37], [135]–[152]. In [30]–[32], the Gaussian prior for target speech is modified to a super-Gaussian prior for frequency-domain sparsity promotion. This is motivated by the observation that speech signals frequency coefficients follow a heavy-tailed distribution [153]. A more expressive, low-rank NMF prior is proposed to model the anechoic speech in [33]–[35]. In [34], [35] however, a different signal model is used, yielding a multi-channel Wiener filter instead of the WPE form. Wang et al. [154] combine both models to propose a super-Gaussian anechoic speech prior with a time-varying scale parameter modelled with NMF. Yoshioka et al. [135] derive an extension of the WPE algorithm considering a Gaussian prior with a full, non-diagonal spatial covariance matrix. Data-based priors can also be considered instead of traditional distributions [P2], [155]–[160], this will be further detailed in Section 1.4.3.

An alternative decomposition of the WPE Filter using Kroenecker products is proposed in [138], yielding higher robustness to observation noise as well as increased computational efficiency. WPE has also been integrated along other multi-channel algorithms like e.g. MVDR or GSC beamforming in [24], [32], [139], [140] for combined denoising and dereverberation, in [141], [142] for simultaneous speaker separation and dereverberation, and in [23], [143] for joint separation, denoising and dereverberation. The integration of WPE and other multi-channel front-end speech enhancement algorithms with automatic speech recognition (ASR) is exhaustively detailed in [7]–[9].

Finally, prior knowledge about the parameters θ of the conditional distribution $p(\mathbf{Y}|\theta)$ can also be included in the optimization via a parameter prior $p(\theta)$. This yields the following maximum a posteriori (MAP) objective:

$$\theta^* = \arg\min_\theta \; -\log\left(\, p(\mathbf{Y}|\theta)p(\theta)\,\right)\,, \tag{1.39}$$

Ito et al. [38] follow such a strategy and propose to add a Gaussian prior with diffuse spatial covariance on the reverberant component. In [39], reverberation is not modelled but the authors instead consider Gaussian microphone-independent additive sensor noise.

Online Multi-Channel Dereverberation

All the abovementioned techniques perform offline processing and are thus not suited for real-time scenarios as in, e.g. hearing devices. Therefore, researchers have dedicated significant effort to finding real-time capable extensions of, e.g. WPE. A first attempt in [147] is based on weighted recursive least squares (RLS) [161, chapter 9], which allows to tackle changing speaker positions by tracking the speaker statistics along time. This single-channel formulation is extended to multi-channel settings in [146]. The optimization objective is a moving average of the original negative log-likelihood function (1.32) with a forgetting factor $\alpha \in]0,1[$. The filter update (1.35) is consequently modified to account for time-dependent filter coefficients $\tilde{\mathcal{G}}(k,f)$:

$$\tilde{\mathcal{G}}(k,f) = \tilde{\mathcal{G}}(k-1,f) + \mathcal{K}(k,f)\left(\mathbf{Y}(k,f) - \tilde{\mathcal{G}}^H(k-1,f)\mathcal{Y}_\Delta(k,f)\right)^H\,, \tag{1.40}$$

where the *Kalman gain* $\mathcal{K}(k,f) \in \mathbb{C}^{MP\times P}$ controlling the speed and the direction of the filter adaptation is given by:

$$\mathcal{K}(k,f) = \frac{(1-\alpha)\mathcal{P}^{-1}(k-1,f)\mathcal{Y}_{\Delta}(k,f)}{\alpha\sigma_d(k,f)^2 + (1-\alpha)\mathcal{Y}_{\Delta}^H(k,f)\mathcal{P}^{-1}(k-1,f)\mathcal{Y}_{\Delta}(k,f)}\,. \tag{1.41}$$

Leveraging the Woodbury identity [162] enables to recursively average the inverse of the correlation matrix $\mathcal{P}$, yielding a computationally efficient estimation of $\mathcal{P}^{-1}(k,f)$ which does not require to invert the updated matrix at each step:

$$\mathcal{P}^{-1}(k,f) = \frac{1}{\alpha}\left(\mathcal{P}^{-1}(k-1,f) - \mathcal{K}(k,f)\mathcal{Y}_{\Delta}^T(k,f)\mathcal{P}^{-1}(k-1,f)\right)\,. \tag{1.42}$$

Kim et al. [145] extend the multi-channel RLS-WPE algorithm described above to the case where the Gaussian prior (1.31) uses a full spatial covariance matrix.

The dereverberation capacity of WPE increases when more microphones and a higher linear prediction are used. However, the correlation matrix $\mathcal{P}$ conversely becomes exceedingly large and its inverse recursive version updated through (1.42) is no longer guarantee to be positive definite, leading to numerical instabilities [161]. To this end, Wung et al. [137] propose a Householder formulation of the RLS-based WPE to increase the stability of online speech dereverberation. As RLS-affiliated techniques recursively average signal statistics, a trade-off between stability and tracking speed is inevitable. Low values of the forgetting factor α accelerate the adaptation mechanism but make the algorithm less stable and more prone to fluctuations and estimation errors. On the other hand, large values of α stabilize inference but over-smoothe signal statistics. To avoid such compromise, RLS-based algorithms can be extended to the celebrated Kalman filter [161], [163], which boasts faster adaptation to, e.g. dynamic acoustic scenarios, while keeping stability at a reasonable level. A first variant of WPE based on Kalman filtering is introduced in [151]. In [150], the state noise second-order statistics used for Kalman filtering are simplified such that the resulting computational complexity is reduced from a quadratic cost to a linear cost with respect to the filter length. Kalman filtering variants of WPE are integrated with, e.g. beamforming in [149], [152] for joint denoising and dereverberation.

Some other online dereverberation techniques leverage a different reverberant signal model from the auto-regressive model (1.29). In [36] the RIRs are considered time-varying and their evolution is modelled using a Markov chain, i.e. a stochastic process which depends only on the previous time increment. Schwartz et al. [37] consider the RIRs to be deterministic and time-varying, and subsequently propose a Kalman filtering approach based on a CTF moving average model of reverberant speech (1.26). In a follow-up work, the same authors directly define the reverberant component and early reflections ATF as the hidden data vector to estimate [164]. The outcome method further accelerates the adaption speed of the Kalman filter to dynamic acoustic scenarios.

1.4 Machine Learning for Speech Dereverberation

Although traditional signal processing techniques have been instrumental in the development of successful dereverberation algorithms, they often rely on distributional assumptions and signal model approximations that are not guaranteed to match real-life scenarios. Instead, machine learning dereverberation algorithms capture the signal properties and structures from

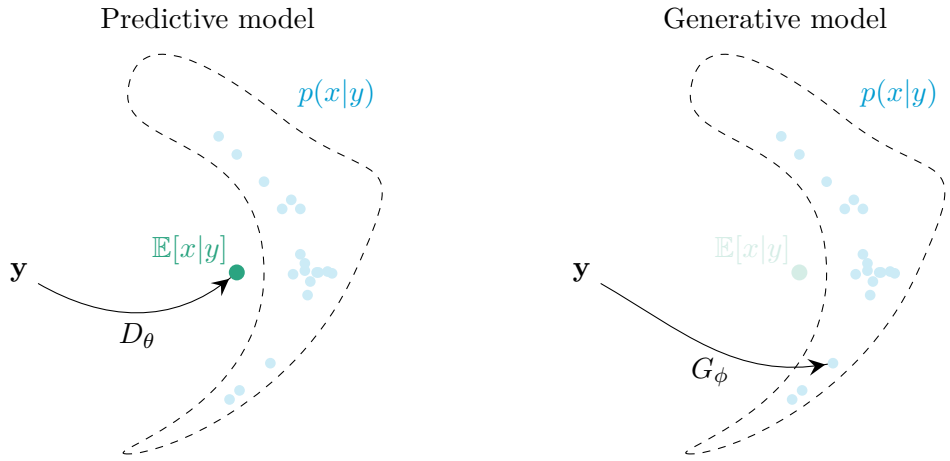


Figure 1.5: *Visualization of the inference process for predictive, conditional generative and hybrid models given a complex arbitrary posterior distribution. The supervised predictive model (left) trained with a least-squares objective regresses to the posterior mean* $\mathbb{E}[x|y]$*. A generative model (center) samples within the posterior distribution* $p(x|y)$*.*

data in order to retrieve anechoic speech. The emergence of DNNs in the past two decades has contributed to a considerable boost in the diversity of data-driven speech dereverberation algorithms. Two complementary categorization systems are generally considered to discriminate between data-driven machine learning techniques.

A first distinction comes between *supervised* and *unsupervised* approaches. *Supervised* models are trained on paired data, in our case anechoic and reverberant speech. Their training objective is to capture the relationships between the reverberant and anechoic speech in order to predict the anechoic speech from the reverberant input at test time. Given their remarkable performance and straight-forward formulations, DNN-powered supervised approaches are prominent in speech restoration [41]. Their main weakness is their lack of robustness to conditions not seen during training. Regardless of their particular internal mechanism, supervised models are trained in a limited-data regime, as it is nearly impossible for a training dataset to faithfully represent all pairs of anechoic and reverberant speech utterances from the real world. Therefore, when supervised models are presented at test time with a reverberant input that significantly deviates from the examples seen during training, the quality of the dereverberated speech is significantly degraded.

In contrast, *unsupervised* models do not need paired data but instead rely only on unlabeled data. We will consider here the most commonly encountered scenario where the training dataset contains only anechoic speech. Unsupervised learning seeks to capture the structure of the presented dataset, and relies on external conditioning at test time to align the model's output with a given reverberant speech utterance. The advantages of unsupervised modelling are two-fold. First, since the task-specific information (i.e. reverberant speech) is injected only during inference, unsupervised methods can leverage a foundational model trained on large-scale clean speech data. Such a pre-trained model can then be employed for various restoration tasks without retraining. Secondly, unsupervised algorithms are by design more robust to various conditions than their supervised counterparts, since they do not rely on any conditioning at training time. However, the overall performance of unsupervised models is usually not as impressive as supervised models trained with paired data, since their conditioning design and training procedures are often more challenging.

The second axis of discrimination within data-driven approaches is related to *predictive* versus *generative* modelling [165]. *Predictive* models optimize a regression (if the output space is continuous, e.g. for waveform prediction) or discrimination (if the output space is discrete, e.g. for even classification) objective. Although this objective can be fully unsupervised in principle, very often, it consists in a point-wise distance between the processed reverberant utterance and the anechoic target. Following empirical risk minimization, a predictive model is optimal when it yields the minimal average error over the training dataset. Predictive models produce a single deterministic anechoic speech estimate, for each reverberant input condition. For instance, when trained with a least squares objective, the optimal model outputs the mean of the posterior distribution of anechoic speech given the reverberant measurement (see Figure 1.5, left). Although significant effort is being currently spent in the direction of explainable artificial intelligence (XAI) [166], systems based on predictive models are largely considered to be black boxes. They are therefore hard to diagnose and interpret, therefore fialing to provide adequate protections for physical safety or ethical fairness.

On the other hand, *generative* models follow a different learning objective: they focus on estimating and sampling from a complex and intractable data distribution. For speech dereverberation, this distribution is the full posterior of anechoic speech given the reverberant measurement (see Figure 1.5, right). Modeling this distribution is more challenging than regressing to the mean of that distribution as typical predictive models would do, but it comes with several advantages. To begin with, it provides a natural measure of uncertainty over the produced outputs, which is a step towards explaining the model's decisions and helps detect failure cases. Furthermore, it allows to sample from the posterior distribution and thus to obtain multiple valid estimates rather than the single deterministic output of predictive models [165]. For instance, this can leave the user the final decision of which sample to select. Another key property of generative models observed in recent studies is their larger robustness to unseen conditions, even when trained in a (semi-)supervised fashion [56], [167].

In this section, we first describe predictive dereverberation models, with a focus on supervised approaches. We then shift to generative models, introducing historical generative models and diving more deeply into diffusion-based generative models [53]–[55]. We dedicate the last part of this section to model-based algorithms that combine data-driven methods with statistical assumptions and algorithmic structures usually adopted in traditional signal processing algorithms.

1.4.1 Supervised Predictive Models

Among learning-based speech processing techniques leveraging DNNs, predictive approaches using supervised learning objectives are the most prominent given their conceptual simplicity and relative ease of training [40], [41]. Using a paired training dataset of anechoic and reverberant speech examples $\mathcal{T} = \{(x_i, y_i)\,,\, i \leq |\mathcal{T}|\}$, supervised predictive models optimize an estimator f_θ parameterized by a DNN with parameters θ, according to a regression objective averaged on the whole dataset:

$$\theta^* = \arg\min_{\theta \in \Theta} \frac{1}{|\mathcal{T}|} \sum_{i \leq \mathcal{T}} \mathcal{L}\left(x_i, f_\theta(y_i)\right) \tag{1.43}$$

Applications of supervised predictive models to dereverberation initially employed fully-connected network architectures [168], then rapidly shifted toward long short-term memory (LSTM) [101] networks given their natural applicability to model speech sequences [169]–[172].

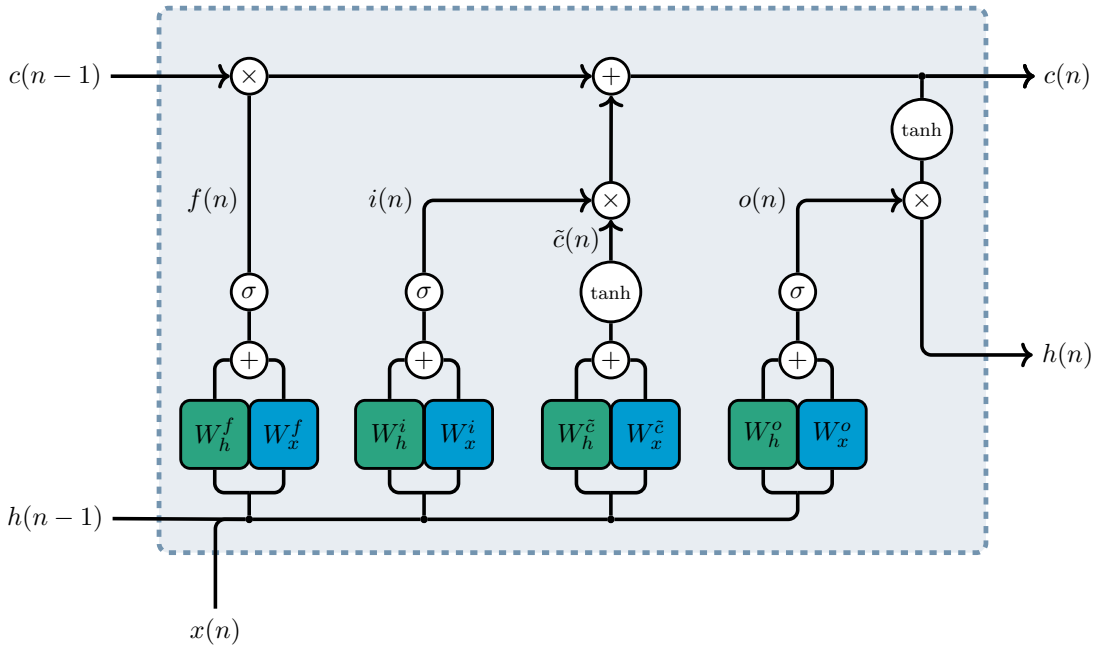


Figure 1.6: *LSTM cell structure [101]. $h(n)$, $c(n)$, $x(n)$ are respectively the hidden state, cell state and input. The parameters learnt during training are $\{W_h^\nu, W_x^\nu\}$ where $\nu \in \{f, i, \tilde{c}, o\}$ represents each of the forget, input, control and output gates (from left to right).*

LSTMs denote a particular flavour of recurrent neural networks (RNNs) [173] that mitigate the vanishing gradient problem encountered in traditional RNNs [174]. An illustration of the structure of an LSTM cell is shown in Figure 1.6. The information flow is controlled by several *gates* with learnable parameters. The input and control gates add information from the cell's input and the past hidden state to the current cell state. The forget gate lets the input and hidden state control how much of the current cell information should be discarded. The output gate then updates the value of the hidden and cell state. Training error is registered in the cell state during backpropagation, which avoids having gradients vanish. The parameters of the gating mechanisms learn to discard or keep the relevant information in the cell state in order to reduce the overall training error.

The early works mentioned above focused on mel-frequency cepstral coefficients (MFCC) prediction since these features were prominent in ASR. Subsequent studies have proposed to work in the STFT time-frequency domain for speech dereverberation, either based on masking [42] or direct mapping [43]–[45]. A phase-aware time-frequency dereverberation algorithm based on the U-Net architecture [175] is introduced in [176]. Many works also investigate the effect of the resolution of the time-frequency representation on dereverberation performance. A convolutional recurrent neural network was proposed in [177] in order to enhance STFT spectrograms obtained computed with various temporal resolutions, and in [178], a neural network combining full- and sub-band frequency processing was adapted for dereverberation. Following the rise of convolutional neural networks (CNNs) in imaging and speech tasks [179], [180], many works have introduced CNN-based methods performing speech dereverberation on the time-domain waveform, where an intermediate representation is learnt by convolutional encoders [47]–[49]. Kothapally et al. [181] propose to use a deformable convolutional network adapting its receptive field based on the detected amount of speech distortion for dereverberation. The most recent success in sequence modeling is the introduction

of the Transformer architecture [103]. Consequently, DNN-based speech dereverberation techniques have rapidly integrated attention mechanisms [46], [182]–[184].

Visual inputs also help humans better understand their reverberant environment, i.e. identifying which materials compose the walls, what is the volume of the room, etc. Some recent works exploit these visual cues and propose audio-visual speech dereverberation [185]–[187]. Furthermore, reverberation being a spatial phenomenon by essence, some works leverage multi-microphone setups for, e.g. multi-channel spectral mapping [50]–[52]. Multi-channel dereverberation can also be performed jointly with downstream tasks such as, e.g. ASR [188]–[191], source separation [192], [193] or denoising [157], [194] through end-to-end training.

1.4.2 Generative Models

Nota Bene

In the rest of this section, we represent signals as being vectors in $\mathbb{R}^D$ where the data dimensionality D includes the time dimension along with the possible extra dimensions.

Although the supervised predictive methods described above are largely prominent for speech processing tasks other than text-based synthesis, recent significant progress in generative modelling has lead to the introduction of generative approaches for dereverberation. Generative models for dereverberation follow a different learning objective compared to their supervised predictive counterparts. Instead of minimizing an objective function between the processed reverberant utterance and the anechoic target (1.43), they learn to estimate and sample from the posterior distribution $p(x|y)$ of anechoic speech given the reverberant speech [P5]. This paradigm can be used to allow the generation of multiple valid estimates instead of a single best estimate as in predictive approaches [165]. Generative modelling also naturally enables to infer a measure of uncertainty regarding the estimated anechoic speech by, e.g. naively measuring the empirical variance of a set of generated samples. Furthermore, it is sometimes possible to incorporate prior knowledge about speech and reverberation signals into generative models. This can help guide the learning process in order to enforce desired properties about the learnt posterior distribution [195], making these methods often more easily interpretable than their predictive counterparts.

Historical Generative Models

Most generative modelling techniques are *hidden variable models*. These models rely on the generic assumption that the data $x \in \mathbb{R}^D$ is generated by a random process involving unobserved variables $z \in \mathbb{R}^{\tilde{D}}$ with prior $p(z)$, such that the data distribution can be expressed as:

$$p(x) = \sum_z p(x, z) = \sum_z p(z)p(x|z)\,, \tag{1.44}$$

where $p(x|z)$ is the data likelihood under the chosen hidden variable model, and the hidden dimension $\tilde{D}$ is in general (but not necessarily) smaller than the data dimensionality D. A simple example of such hidden variable model is the Gaussian mixture model (GMM), which considers that data is generated by a mixture of C Gaussian sources with mixing weights $\{\pi_n \in]0, 1[\}_{c=1}^{C}$. In this case, the hidden variable prior is a categorical function

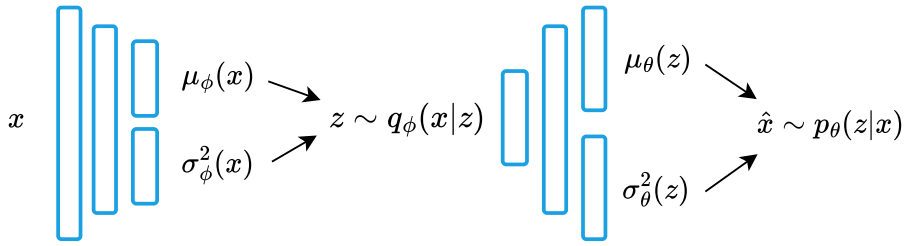


Figure 1.7: *Structure of a variational auto-encoder [197]. The decoder parameters θ control the generative model $p_\theta(z|x)$, i.e. the construction of clean speech x given a Gaussian hidden variable z. The hidden variable space is itself learnt through a variational approximation with parameters ϕ of the intractable posterior distribution $q_\phi(z|x) \approx p(x|z)$.*

$p(z) = \text{Cat}(z_1, \ldots, z_C)$ and the data distribution is parameterized as:

$$p(x) = \sum_{c=1}^{C} \pi_c \mathcal{N}(\mu_c, \sigma_c^2)\,. \tag{1.45}$$

An early generative model for speech denoising and dereverberation is presented in [196], where a frequency-domain Gaussian mixture prior is trained on a dataset on clean speech. At test time, the parameters of a probabilistic posterior model for noisy speech are inferred through a variational EM algorithm.

GMMs are *shallow generative models*, i.e. there is a finite number of hidden variables and the likelihood function has a very simple form. Instead, *deep generative models* employ one or more layers of continuous hidden variables, generally leveraging DNNs for their remarkable abilities to learn complex representations of data [165]. In this category, variational auto-encoders (VAEs) [197] perform explicit density estimation. They posit that the observation model follows a Gaussian distribution parameterized by an decoding neural network with parameters θ and that the hidden variable itself follows a standard Gaussian distribution:

$$p(z) = \mathcal{N}(0, \mathcal{I}) \;\text{ and }\; p_\theta(x|z) = \mathcal{N}(\mu_\theta(z), \sigma^2_\theta(z))\,, \tag{1.46}$$

where $\mathcal{I}$ is the identity matrix. A VAE can be understood as a GMM with an infinite number of components N. Training a VAE requires access to the true posterior distribution $p_\theta(z|x)$, which is intractable. Therefore a Gaussian variational approximation $q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \sigma^2_\phi(x)) \approx p_\theta(z|x)$ is used instead, whose mean and variance are provided by an encoding network with parameters ϕ. The encoder parameters ϕ and decoder parameters θ are jointly trained using a variational lower bound on the intractable likelihood:

$$\log p_\theta(x) \geq \mathbb{E}_{z\sim q_\phi(z|x)} \left[\log p_\theta(x|z)\right] - \text{KL}\left[q_\phi(z|x)||p(z)\right]\,. \tag{1.47}$$

The first term in (1.47) is the reconstruction error and the second term regularizes the hidden variable space to ensure that it encodes meaningful information, The structure of a VAE is summarized on Figure 1.7.

VAEs have been used for unsupervised single-channel dereverberation [198], [199]. In these works, the generation of anechoic speech is conditioned on a reverberant recording by optimizing a ML objective. The optimization is based on a Monte Carlo Expectation Maximization (MCEM) algorithm [200], which is a stochastic variant of the EM algorithm using a Monte-

Carlo approximation of the intractable posterior sampling update [201]. Wang et al. [199] extend the method in [198] by using a recurrent VAE, where the hidden variable is conditioned on past hidden variables and input speech values.

Generative adversarial networks (GANs) [202] do not actually belong to the category of hidden variable models but instead perform implicit density estimation. They train a DNN called *generator* to directly produce samples which are supposed to be indistinguishable from real data. The sample output by the generator is processed by another DNN denoted as *discriminator*. This latter assesses the likelihood of the presented sample, i.e. it learns to discriminate whether the produced sample was produced by the generator or actually belongs to the data distribution. The generator and discriminator are trained in an adversarial fashion, playing a min-max game to ultimately increase the capacity of the generator to output samples likely to belong to the target data distribution. The generation of anechoic speech is guided by a reverberant conditioning signal, such that the generated sample belongs to the posterior distribution given the reverberant measurement [203], [204]. This is applied for dereverberation in [183], [205], [206]. In [207], an unsupervised framework is proposed in order to train on unpaired data. However, an important caveat must be mentioned here. The aforementioned GAN-based dereverberation techniques exclude the latent Gaussian sample altogether, and therefore they cannot generate multiple anechoic estimates but only learn a deterministic mapping. This disqualifies them as generative models, according to our definition that we borrow to Murphy et al. [165].

GANs and VAEs have been key in the early developments of generative models but have their own downsides. They both have the tendency of experiencing mode collapse, which occurs when the input noise sample is ignored and the model always returns samples belonging to one of the modes of the learnt distribution. Model collapse results in a low expressivity of the generative model, since only a couple modes of the target distribution are learnt, instead of the complete distribution. Furthermore, GANs are hard to train because of the adversarial mechanism, while VAEs generally lack expressivity because of their simplistic architecture and distributional assumption.

Diffusion Models

More recently, *diffusion models* [53], [55] have been introduced as a class of deep generative models solving the aforementioned issues. The key mechanism of diffusion models is that they break down the problem of generating complex, structured, high-dimensional data into a series of easier *denoising* tasks. A DNN is trained to address the task of removing Gaussian noise from corrupted samples, such that new data samples can be generated by iteratively denoising an initial Gaussian sample. Under certain conditions detailed hereafter, the hidden space explaining the data is a continuum of Gaussian variables, motivating the comparison of diffusion models to infinitely deep VAEs [208]. It has been rapidly observed that diffusion models are easier to train than GANs and much more expressive than VAEs, yielding state-of-the-art quality performance for density estimation of, e.g. natural images [53], [55], [209], music [210], and speech [211], [212].

We first define the *forward diffusion process*, indexed by the continuous variable $\tau \in [0, T]$ called *process time*.

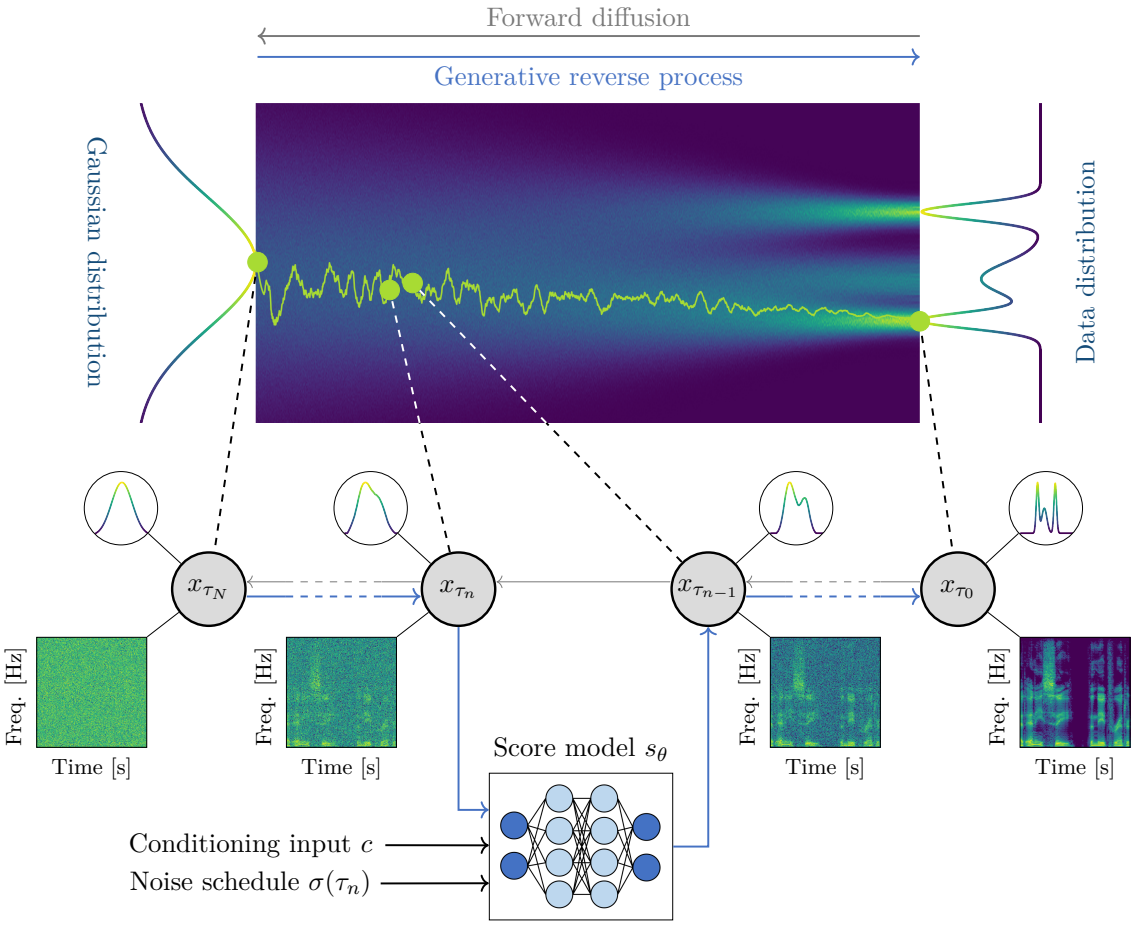


Figure 1.8: *A continuous-time diffusion model [55] transforms (left) a Gaussian distribution to (right) an intractable data distribution through a stochastic process* $\{x_\tau\}_{\tau\in[0,T]}$. *The forward diffusion process* (1.48) *(right to left) transforms the data and adds Gaussian noise. The reverse diffusion process* (1.50) *(left to right) generates data by iteratively denoising the Gaussian sample using the score model* s_θ.

Nota Bene

The process time must not be mistaken with the signal time dimension indexed with the variable n in the previous sections.

Stochastic processes are indexed with the process time τ as a subscript (ex: $u_\tau \in \mathbb{R}^D$), while regular functions of the variable τ use the argument in parentheses (ex. $u(\tau) : [0, T] \to \mathbb{R}$). This notational convention makes the following more consistent with the publications in chapters 4 and 5.

This stochastic process maps a clean speech utterance $x \in \mathbb{R}^D$ from the target distribution to a Gaussian sample $x_T \in \mathbb{R}^D$ by iteratively adding Gaussian noise and rescaling the data. The forward process can be obtained by solving the following stochastic differential equation (SDE) with values in $\mathbb{R}^D$:

$$\mathrm{d}x_\tau = f(x_\tau, \tau)\,\mathrm{d}\tau + g(\tau)\,\mathrm{d}w_\tau\,, \tag{1.48}$$

where the function $f : \mathbb{R}^D \times \mathbb{R} \to \mathbb{R}^D$ is referred to as the *drift coefficient* and relates to the

deterministic part of the SDE. The function $g : \mathbb{R} \to \mathbb{R}$ is called the *diffusion coefficient* and controls the amount of randomness in the SDE. More precisely, the diffusion coefficient g scales the noise injected by the stochastic process $w_\tau \in \mathbb{R}^D$. In most cases, w_τ is chosen to be a *Wiener process*, which is a stochastic process with independent and normally distributed increments, i.e. $w_{\tau+\mathrm{d}\tau} - w_\tau \sim \mathcal{N}(0, \mathrm{d}\tau\, \mathcal{I})$ [213]. If the SDE drift f is an affine function of x_τ and the diffusion g is independent of x_τ, then the transition kernel has a simple Gaussian form [213]:

$$q_\tau(x_\tau|x) = \mathcal{N}(\mu(x,\tau), \sigma^2(\tau) I)\,, \tag{1.49}$$

where the mean $\mu(x,\tau)$ and standard deviation $\sigma(\tau)$ can be obtained by solving the SDE and computing the first and second moments of the solution [213].

New speech samples can be generated by the reverse diffusion process, which is a stochastic process parameterized by a process time flowing in the reverse direction compared to the forward process. Under mild regularity conditions with respect to the SDE coefficients f and g, reversing the process time axis in (1.48) results in another SDE called the *reverse SDE* [214]. Notably, the marginal distributions of the reverse process are also Gaussian and match those of the forward SDE. Therefore, generating a new sample x boils down to solving the reverse SDE:

$$\mathrm{d}x_\tau = \Big[f(x_\tau,\tau) - g^2(\tau) \nabla_{x_\tau} \log p(x_\tau) \Big]\, \mathrm{d}\tau + g(\tau)\, \mathrm{d}\bar{w}_\tau\,, \tag{1.50}$$

where $\mathrm{d}\tau < 0$ as the process axis is traveled in the reverse direction. The stochastic process $\bar{w}_\tau \in \mathbb{R}^D$ is another Wiener process associated to this reverse process axis. The *score function* $\nabla_{x_\tau} \log p(x_\tau) \in \mathbb{R}^D$ is generally intractable and is approximated with a DNN called *score model* $s_\theta(x_\tau, \tau)$ with parameters θ. Directly trying to learn the true score function $\nabla_{x_\tau} \log p(x_\tau)$ is computationally challenging and runs into two pitfalls [54]. First, most of the complex empirical data distributions lie on a low-dimensional manifold (which is referred to as the *manifold hypothesis*) and the score function is ill-defined on such manifold. Second, regions of low probability density are not well represented in the training dataset, thus the score model struggles to learn the score function in such low-probability regions [54, Figure 2]. Therefore in practice the parameters of the score model are often optimized using the *denoising score matching* objective [215], i.e. matching the score of the Gaussian transition kernel $q_\tau(x_\tau|x)$:

$$\mathop{\mathbb{E}}_{\substack{x\sim p(x) \\ \tau\sim\mathcal{U}(0,T) \\ x_\tau\sim q_\tau(x_\tau|x)}} \left[\lambda(\tau) \left\| s_\theta(x_\tau,\tau) + \frac{x_\tau - \mu(x,\tau)}{\sigma(\tau)^2} \right\|_2^2 \right]. \tag{1.51}$$

The denoising score matching objective (1.51) can be described as follows. First, a data example x is sampled from the training set. Then, a process time τ is sampled uniformly between 0 and T, and the diffusion state x_τ is obtained by sampling from the transition kernel (1.49), resulting in an affine transformation and additive Gaussian noise. The score model $s_\theta(x_\tau, \tau)$ is then trained to learn the score of the Gaussian kernel, i.e. to estimate how much Gaussian noise was added and how to invert the affine transformation. Adding Gaussian noise allows to escape the low-dimensional manifold and to have well-defined gradients in all regions, and also helps with learning the score function in low-probability regions. The process time dependent scaling factor $\lambda(\tau)$ is chosen empirically in order to stabilize training [55], [216]. Once the score model $s_\theta(x_\tau, \tau)$ has been trained, the reverse SDE (1.50) can be solved with a numerical solver [217]. First the process time axis must be discretized into N steps $\{\tau_N, \tau_{N-1}, \ldots, \tau_0\}$ with a step size $\Delta\tau_n := \tau_n - \tau_{n-1}$, often chosen as uniform. The initial condition x_{τ_N} can be sampled either as close as possible to the final state in the forward

process, or it can include additional cues about the clean speech to guide the generation process, as in, e.g. [P8], [218].

Several speech enhancement and dereverberation systems have been designed using supervised diffusion models. Lu et al. [219] proposed a first waveform-based model which conditions the generation of clean speech by a noisy speech utterance passed as auxiliary input to the score model. In [P7], [P8], [56], [220], [221] instead, the conditioning is directly included in the diffusion process. The drift coefficient is modified such that the mean of the Gaussian transition kernel interpolates between degraded and clean speech. Later works [222], [223] apply this principle to multi-channel settings. In [223] the conditioning of the diffusion process is expanded with signals processed by traditional multi-channel enhancement schemes such as, e.g. WPE [224] and MVDR beamforming. Unsupervised dereverberation using diffusion models is investigated in [P10], [P12], [P11]. A first informed method uses the knowledge of the RIR to perform deconvolution of the reverberant signal model, with a diffusion-based prior on anechoic speech [P10]. The proposed algorithm is extended to the blind case in [P12], [P11] where the RIR is estimated using a statistically-informed subband filter.

Although they have not been yet used for speech dereverberation, to the best of our knowledge, we introduce flow-based models out of completeness of our overview on generative models. Flow-based models [225]–[227] follow the same divide-and-conquer strategy as diffusion models. Normalizing flows, either discrete [226] or continuous [225], formulate density estimation as a series of invertible mappings between the target distribution and a simple prior, e.g. a standard Gaussian. The invertibility of the intermediate mappings guarantees that the learnt probability distribution is normalized, hence the name of this method. Normalizing flows have been used for speech enhancement in [228] in a supervised fashion, and in conjunction with VAEs in [229] for unsupervised multi-channel speech enhancement using a NMF-based noise model. Flow matching models [227] provide a different training objective for continuous normalizing flows, drawing a connection to the diffusion paradigm. Flow matching generalizes Gaussian diffusion models as it allows for more flexible parameterization of the probability paths between the initial and target distributions, based on, e.g. optimal transport [230]. Furthermore, the actual choice of the terminating distributions in flow matching models is not restricted to Gaussians, as is the case in diffusion models. Liu et al. [167] pre-train a foundational speech model with unsupervised flow matching. The resulting model can be finetuned to perform several speech-related tasks, including speech enhancement. Moliner et al. [231] train a conditional flow matching model to perform acoustic transfer between two reverberant conditions using unpaired data. This approach can actually be used for dereverberation if the target distribution is chosen to represent an anechoic setting.

The main drawback of the diffusion (or flow matching) framework is the iterative nature of the reverse process for inference. Numerical solvers provide polynomial approximations of the true integral solution to the reverse SDE (1.50) [217]. The error between the true solution and the approximation computed by the numerical solver over the full trajectory is called *global truncation error*, and is of magnitude $\mathcal{O}(\frac{1}{N^r})$, where r is the *order* of the solver. In most cases, the order r is also the degree of the polynomial function used to approximate the true process, and therefore the score model is evaluated r times at each step of the reverse process. This yields a number of DNN evaluations of magnitude $\mathcal{O}(Nr)$. Therefore, using more diffusion steps and a higher-order solver reduces the global truncation error but increases the computational cost. Solutions to avoid a harsh trade-off between generation quality and computational cost include approximating the reverse process schedule [232], [233] distilling the reverse process [234], [235], optimizing the reverse noise variance schedule [236],

[237] or using consistency models [238]–[240]. Some solutions are examined in [P5, Section "Practical Requirements of Diffusion-based Sampling for Audio Tasks"] in the context of audio restoration tasks.

1.4.3 Model-based Dereverberation

Model-based algorithms are hybrid methods combining DNN-based estimators with elements usually found in traditional signal processing algorithms. Such elements can consist of prior knowledge of room acoustics, algorithmic structures, or distributional assumptions. Good model-based methods benefit from the best of both worlds. On the one hand, they perform better than traditional algorithms thanks to the powerful non-linear regression and density estimation abilities of neural networks. On the other hand, leveraging algorithmic structure from successful traditional approaches and injecting prior knowledge help mitigate the flaws of DNN-based techniques with respect to, e.g. computational complexity or generalizibility.

For instance, [241] employs a DNN for estimation of source dominance in a ML framework. Several works estimate the variance of anechoic speech with DNNs to accelerate the convergence of WPE [P2], [155]–[159]. In [155], a fully-connected network is used for block-online DNN-assisted WPE. Heymann et al. [156] adapt this work for online RLS-based WPE processing using LSTMs. We propose in [P2] to integrate the RLS-WPE computations into the optimization graph of the anechoic speech DNN estimator for end-to-end learning. Petkov et al. [158] design an unsupervised framework for optimizing a DNN-assisted WPE algorithm, trained on reverberant speech only. In [160], an unsupervised speech denoiser is utilized as a plug-and-play prior for anechoic speech, extending the WPE maximum-likelihood objective with the regularization by denoising (RED) criterion [242]. These DNN-guided WPE algorithms can be further combined with other DNN-assisted traditional processing such as, e.g. MVDR beamforming [157], [190], [243], single-channel post-filtering [P1], principal component analysis [159] or acoustic echo cancellation techniques [194]. In a closely related fashion, [223] propose to leverage WPE and beamforming estimates as auxiliary inputs to a multi-channel conditional diffusion model. The authors show that the WPE-processed conditioning inputs help improve the robustness of the diffusion model to new acoustic and noise conditions.

Model-based methods can also leverage reverberant signal models or room acoustic characteristics as domain knowledge for DNN-based dereverberation. Wang et al. [193] propose to use the CTF assumption (1.4) to derive auxiliary losses for multi-channel speech dereverberation. In [198], [199], the CTF approximation is also used, this time to provide an observation model for MAP-based unsupervised speech dereverbation using a VAE prior. A similar approach is taken in [P12], [P11], where the CTF observation model is combined with a diffusion-based prior. In [44] the T_{60} reverberation time is provided as an auxiliary input to inform the model about the acoustics of the room. Barhman et al. [45] constrain their speech dereverberation estimation on EDRs estimated given a physical model of reverberation learnt jointly with the speech model. Finally, we propose in [P4] a simple DNN-based technique for extending mask-based processing to deep filtering for dereverberation. This approach is motivated by the intuition that deep filtering can serve as an approximate inverse to the CTF model.

1.5 Room Impulse Response Generation and Room Acoustics Estimation

We start this section by presenting RIR generation techniques using geometrical, stochastic and DNN-based methods. The remainder of the section is dedicated to the estimation of room acoustics, ranging from the prediction of acoustic parameters like the T_{60} to the complete estimation of the RIR.

Room Impulse Response Generation

Simulating a plausible acoustic environment is vital for the perceived realism of a virtual scene, especially given a visual input as in virtual reality (VR) scenarios. Other applications like, e.g. music production, can also benefit from highly realistic room simulation.

Furthermore, having access to a large number of RIRs is the most straight-forward strategy for generating paired anechoic and reverberant data. Such datasets are used to design supervised training pipelines, not only for training dereverberation algorithms but also as a solution for making DNNs tackling tasks like speech restoration, recognition or localization generalize well to real-life scenarios where reverberation plays a substantial role [244]. The latter is a form of *data augmentation* [245], which denotes a transformation of the original data to increase the diversity of training conditions. However, collecting real recordings of RIRs is a time- and resource-consuming activity. In contrast, RIR simulation enables to generate paired anechoic and reverberant datasets in an efficient and affordable fashion.

RIR simulation techniques should ideally have a fine-grained control over a variety of acoustic parameters. This includes frequency-dependent T_{60} profile, microphone-speaker distance, source directivity, reverberant field spatial coherence, etc. Furthermore, Srivastava et al. [246] experimentally demonstrate that it is worthwhile increasing the realism of RIR simulation, for instance by considering directional sources and receivers as well as frequency-dependent wall absorption profiles. They demonstrate that doing so significantly improves the performance of DNN-based acoustic parameter estimators trained on simulated data only, when tested on real reverberant utterances.

As room acoustics behave differently in various frequency regions, there is no RIR simulation technique perfectly covering the whole frequency domain. At low frequencies, i.e. below the Schroeder frequency f_g (1.12), resonant modes play an important role in room acoustics and therefore RIR models rely mostly on harmonic solutions of the wave equation obtained via numerical solvers [250]. In the medium frequency range, typically located between f_g and $4f_g$, statistical techniques are generally used. For instance, Polack [60] models the time-domain RIR as a stationary Gaussian process with exponential amplitude decay (1.11). The frequency-domain ATF of the reverberant component may also be modelled as a stochastic process with frequency-dependent energy spectrum and spatial coherence [59].

In frequency regions above $4f_g$, room acoustics are reasonably well represented by geometrical acoustics. Interestingly, this covers a wide audio frequency range for most rooms. If we consider our previous example in Section 1.2.2 of a room with dimensions $8 \times 6 \times 3$ meters and $T_{60} = 0.4$ s, geometrical acoustics would apply for all frequencies above $4f_g \approx 420$ Hz. This includes a large part of the human speech frequency spectrum. Geometrical methods for RIR simulation include ray-tracing [251], [252] and the image source method (ISM) [248]. Ray-tracing follows phonons, i.e. virtual sound particles, in order to simulate the effect of acoustic waves when they get reflected across the room. ISM-based methods consider instead

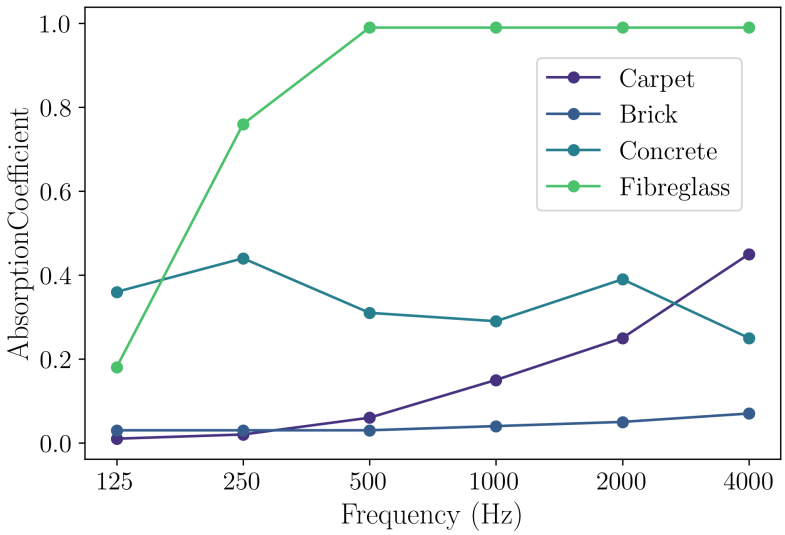
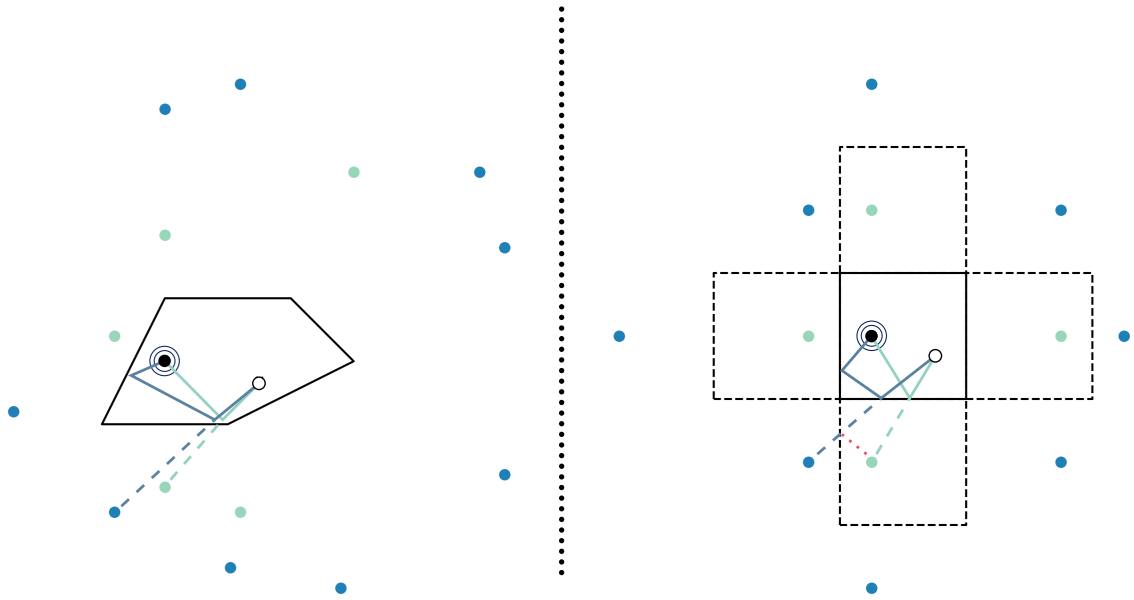


Figure 1.9: *Absorption profiles for different materials. Data from [247].*



(a) *Complex geometry* (b) *Shoe-box geometry*

Figure 1.10: *Image Source Model [248]. Virtual sources corresponding to first- and second-order reflections are shown in light green and blue, respectively. The original source is the large black dot, and the receiver is the black-circled white dot. A paramount property of the shoebox geometry is that $(n+1)$-th order virtual sources are themselves first-order virtual sources of n-th virtual sources (see the dashed red line on Figure 1.10b). This drastically simplifies and accelerates the computation of the image source model for larger reflection orders in shoebox rooms. Figure plotted using* `pyroomacoustics` *[249].*

that reflections can be seen as originating from virtual sources located beyond the walls of the room, as pictured on Figure 1.10. This technique is used in numerous simulation engines given its conceptual simplicity and ease of implementation. However, there are many known limitations of ISM-based simulation. ISM is impractical for representing long impulse responses since the method's computational complexity is exponential with respect to the order of the reflections. Therefore in practice, most RIR simulation engines like, e.g. `pyroomacoustics` [249] only model a handful of these early reflections with ISM, leaving the modeling of the reverberation tail to statistical methods. Furthermore, high-order image sources are much easier to locate for simple room geometries such as, e.g. shoebox-shaped rooms, which excludes numerous real-life scenarios (see Figure 1.10). Tang et al. [253] propose to approximately model occlusion and diffuse reflections via a Monte Carlo estimation of the ray-tracing method. Recently, a RIR simulation method combining ISM, diffuse acoustic modeling as well as approximation of the wave equation has been introduced in [254]. A full stochastic approximation of ISM is proposed in [255] for fast RIR generation. However the resulting acoustics are not geometry-dependent and very little control is given on the conditioning of the generated RIRs. This method is therefore fit for data augmentation but not for simulating a specific environment for, e.g. VR applications.

DNNs have also been recently used for simulating and generating RIRs. Ratnarajah et al. [256] propose FAST-RIR, an algorithm based on conditional GANs for generation of single-channel RIRs in shoebox-shaped rooms. The authors condition the simulation on room parameters such as the room dimensions, listener and speaker positions as well as T_{60} time. As mentioned in Section 1.4.2, the conditional glsgan used in this work is actually not generative in the sense that it only can predict a single estimate given the acoustic conditions, since the authors do not use a noise vector at the input of the generator. A conditional diffusion model is proposed in [257] for multi-channel RIRs generation. The model is explicitly conditioned on the relative microphone-speaker position such that the inter-microphone cues of the direct path are respected. The room-dependent conditions are provided to the model via a self-supervised room embedding learnt through contrastive learning [258].

Room Acoustics Estimation

Estimating room acoustics is another long-lasting research field. Room acoustic estimation tasks range from predicting the T_{60} or the room geometry, all the way to replicating the RIR given only a reverberant recording captured in the room.

Estimating the T_{60} given a reverberant recording is a classical task and has a handful of applications. Methods based on ML objectives and statistical approximations like, e.g. Polack's model [60], [259] are presented in [260], [261]. Löllmann et al. [262] provide an overview and comparison of blind T_{60} estimation methods using traditional algorithms and DNNs.

Estimating the room geometry and wall material properties is an arduous task, even when the RIR is available. The absorption profiles of room walls are estimated in [263] from RIRs, considering the room geometry approximately known. The method combines ISM with a probabilistic model of the TDOAs to develop a search algorithm looking for RIR regions where single reflections can be isolated. The absorption profiles are then estimated from these regions using non-convex optimization. Sprunck et al. [264] locates virtual sources in ISM-simulated RIRs, which enables to infer the geometry of the room. They first relax the non-convex optimization problem of recovering continuous spike locations given discrete RIR measurements into a convex optimization problem. The method successfully retrieves numerous virtual source positions, although the considered scenario is relatively artificial

as the authors neglect realistic acoustic behaviours like, e.g. directional sources. Joining the two aforementioned tasks, a full inversion procedure of the ISM is proposed in [265] in order to retrieve the source and receiver positions as well as the room dimensions and wall absorptions.

Given their remarkable abilities for non-linear regression, DNNs are getting more and more involved for room parameter and RIR estimation. A new algorithm for single-channel T_{60} estimation is presented in [266] using CNNs, while Parada et al. propose to estimate the C_{50} clarity index using a bidirectional LSTM operating in the frequency-modulation domain [267]. In [268], the frequency-dependent profile of the T_{60} is estimated using a fully-connected network conditioned by the room geometry. In [269], the authors combine modulation transfer functions, Schroeder's RIR model and CNNs to simultaneously estimate several acoustic parameters including the T_{60}, C_{50} and early decay time. Foy et al. [270] design a DNN-based method for estimating the frequency-dependent absorption profiles of rooms, comparing fully-connected and convolutional architectures. In real situations, source positions are often changing over time, yielding dynamic acoustic scenarios. Götz et al. [271] show that a joint estimation of T_{60} and C_{50} is possible in dynamic scenarios employing convolutional recurrent networks operating in the Gammatone filterbank representation [95].

Full RIR estimation using only reverberant signals is as ill-posed a problem as blind dereverberation given the symmetry of the convolutional signal model (1.2). Several DNNs-based techniques have been proposed to solve this task. The FiNS algorithm [272] uses a predictive RIR estimator parameterized by a DNN divided in two components. The first component is a one-dimensional CNN encoder. The decoder part consists in a module directly estimating the direct path and early reflections, and a noise-filtering stage with trainable finite impulse response (FIR) filters. This noise-filtering module is dedicated to model the late reverberation tail, inspired by the statistical model by Polack [60], [259]. Lee et al. [273] apply FiNS to the task of estimating a room-representative impulse response in the presence of multiple simultaneous sound sources. In [274], they propose a DNN-based RIR estimation technique leveraging differentiable approximations of reverberation models such as filtered velvet noise [275] and delay networks [276]. In [277], the same authors shift toward estimating RIRs using auto-regressive generative modelling. Inspired by Transformer language models, they first train a vector quantized VAE to represent RIRs as discrete tokens. They then cast RIR estimation as a conditional generation task, using the reverberant speech utterance as a conditioning signal to the Transformer model. Finally, we present in [P12], [P11] an unsupervised diffusion-based generative model which performs blind speech dereverberation and RIR estimation in a single-channel setup. The RIR is modelled using a subband filter with exponential decays, extending Polack's statistical model [60], [259] with frequency-dependent decay rates. Although the model is tailored to model the free decay period, leaving the estimation of early reflections up to the free STFT phases, the proposed method captures reasonable acoustic properties of the rooms as shown by the good T_{60} and C_{50} prediction performance on the estimated RIR.

1.6 Outline and Contributions

This thesis collectively treats model-based algorithms and diffusion-based generative models for speech dereverberation. We organize this dissertation as follows.

Model-based Speech Dereverberation

This first chapter investigates model-based techniques for speech dereverberation, combining aspects of traditional signal processing and signal model approximations with deep learning elements.

Research Questions

RQ1 Are there any benefits of integrating traditional algorithmic structures in DNN-based frameworks, compared to using pure neural networks trained end-to-end?

RQ2 How can one integrate the knowledge of the convolutional signal degradation model into the design of DNN-based dereverberation algorithms?

In Section 3.1, we propose a real-time capable two-stage dereverberation algorithm tailored for hearing device users [P1]. The first processing stage extends the frame-online DNN-assisted WPE algorithm in [156] by including the WPE computations into the optimization graph of the LSTM network estimating the anechoic speech variance. This modification shifts the focus of the training objective toward yielding optimal dereverberated speech at the output, rather than optimal intermediate anechoic speech variance. The second stage is a single-channel Wiener filter where the target speech and residual reverberation are estimated by two LSTMs. The first stage removes most of the moderate reverberation accessible within the WPE filter range, thereby benefiting the second stage which only needs to remove residual reverberation statistically uncorrelated with the target speech.

In Section 3.2, we investigate and compare the signal models behind speech denoising and dereverberation [P4]. We show that when considering reverberant signals, a DNN trained to estimate a subband multi-frame filter (often called a *deep filter* [278]) yields better performance than the same DNN trained to perform time-frequency masking (i.e. a subband filter with only one frame). Since the single-to-multi frame extension module has very few parameters, it enhances the dereverberated speech quality with no tangible influence on the computational cost. However, when signals are degraded by additive noise only, the performance does not improve by turning the time-frequency mask into a deep filter. We motivate this result by connecting it to the fact that the signal model for reverberation is well represented by the subband approximation (1.4) whereas that of additive noise perfectly respects the narrowband approximation (1.5).

Supervised Conditional Diffusion Models for Speech Dereverberation

This chapter focuses on dereverberation techniques using diffusion models trained in a supervised fashion. Harnessing conditional generative models moves away from predictive models that learn a regression rule connecting reverberant and anechoic speech, and instead focuses on estimating the conditional distribution of anechoic speech given on a reverberant recording.

Research Questions

RQ3 How do diffusion-based generative models compare to predictive models in terms of speech restoration performance, robustness and generalization to unseen conditions? Is there a dependency on the restoration task at hand?

RQ4 Can one find a suitable combination of predictive and diffusion-based generative models for speech restoration? What are the potential advantages of such hybrid framework compared to pure predictive or generative modeling?

In Section 4.1, we give a tutorial on diffusion models for audio restoration (which includes speech dereverberation) [P5]. There, we introduce basic concepts of diffusion-based generative modeling and present a review of the current state of the art.

Section 4.2 is dedicated to a comparative analysis of predictive methods versus diffusion-based generative models [P7]. Echoing Section 3.2, emphasis is put on the dependency of each model's relative performance with respect to the task at end, which in this study consists in speech denoising, dereverberation and bandwidth extension. We show here that diffusion models consistently outperform their predictive counterparts (using the same neural architecture) in terms of speech quality across all tasks. More interestingly, predictive models seem to perform very well on denoising, but significantly worse than diffusion models when considering non-additive signal degradations such as reverberation and bandwidth reduction. This suggests that better capturing complex posterior distributions (such as, e.g. that of anechoic speech given reverberant speech) helps diffusion-based generative models outperform predictive models which only regress to the mean of these distributions.

This hypothesis is given a closer look in Section 4.3, where we propose to combine predictive and generative modeling to get the best of both approaches [P8]. The rationale at the root of the corresponding publication can be summarized in the following, and is illustrated in Figure 1.5. Estimating the mean of the posterior distribution with a L^2-optimized predictive model should serve as a good cue for modeling the whole posterior distribution. We suggest leveraging a predictive model to provide an intermediate step toward modeling the posterior distribution with a diffusion model. The proposed modification significantly increases the speech enhancement and dereverberation performance of the generative model, while simultaneously reducing the number of iterations needed for reverse diffusion.

Solving Single-Channel Speech Dereverberation as an Inverse Problem with Unsupervised Diffusion Models

The final chapter treats the topic of unsupervised dereverberation with diffusion-based generative approaches. We frame dereverberation as a deconvolution task, viewing it through the lens of inverse problems, similar to traditional informed methods. This section reflects the themes of the first chapter, offering a detailed exploration of how domain knowledge can assist the design of inverse problem-based dereverberation.

Research Questions

RQ5 Can diffusion models provide a good prior for regularizing the inverse problem of single-channel informed dereverberation? What is the resulting robustness with respect to noise in the reverberant recording and errors in the RIR?

RQ6 Can one leverage diffusion models and domain knowledge to jointly estimate the room acoustics and anechoic speech from a single-channel reverberant utterance?

Section 5.1 introduces matters by turning to the easier task of informed dereverberation, i.e. retrieving the anechoic speech when both the reverberant measurement and the RIR are accessible. As mentioned in Section 1.3, informed dereverberation is easier than blind dereverberation but it is nonetheless an ill-posed inverse problem. Therefore, an informative prior on anechoic speech is required for an inverse problem solver to converge to a reasonable solution. We propose to employ a prior based on an unconditional diffusion model, trained on anechoic speech only [P10]. An unsupervised model then combines this prior with a likelihood model of the reverberant signal model via the so-called diffusion posterior sampling (DPS) framework [279]. We demonstrate that the method yields state of the art results for informed dereverberation and high robustness to observation noise.

However, the aforementioned method has two main weaknesses. As all informed methods, it is very sensitive to slight fluctuations in the allegedly known RIR. But most importantly, it treats informed cases, which only depict a minority of actual real-life scenarios. Therefore, in Section 5.2, we extend this method to the blind scenario by designing a RIR estimator based on a subband filter with frequency-dependent exponential decays [P12]. This estimator follows statistical observations from [60], [259] and uses the subband approximation (1.4), therefore qualifying as a model-based algorithm. We show that the proposed method can provide a high-quality anechoic speech estimate compared to previous unsupervised state of the art. Furthermore, we simultaneously obtain an estimate of the RIR that captures room acoustics well, as represented by, e.g. T_{60} and C_{50} frequency profiles. We empirically demonstrate the high robustness of this unsupervised method to various acoustic scenarios. The proposed method comes on par or even outperforms competitive supervised baselines in a mismatched scenario where the test acoustic conditions differ from those the supervised models were trained on.

2

Overview of the Related Publications

The list of publications in this thesis follows the research plan outlined in Section 1.6. Peer-reviewed papers included in the body of this cumulative dissertation are highlighted in a blue box, while all other related publications and pre-prints are left to Appendix A.

Chapter 3. Model-based Speech Dereverberation

3.1. Lightweight Model-Based Dereverberation for Hearing Devices

[P1] J.-M. Lemercier, J. Thiemann, R. Konig, and T. Gerkmann, "A neural network-supported two-stage algorithm for lightweight dereverberation on hearing devices," *EURASIP J. Advances in Signal Process.*, vol. 2023, no. 18, pp. 1–12, 2023.

[P2] J.-M. Lemercier, J. Thiemann, R. Konig, and T. Gerkmann, "Customizable end-to-end optimization of online neural network-supported dereverberation for hearing devices," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022.

[P3] J.-M. Lemercier, J. Thiemann, R. Koning, and T. Gerkmann, "Neural network-augmented Kalman filtering for robust online speech dereverberation in noisy reverberant environments," in *Proc. Interspeech*, 2022.

3.2. Deep Subband Filtering for Speech Dereverberation

[P4] J.-M. Lemercier, J. Tobergte, and T. Gerkmann, "Extending DNN-based multiplicative masking to deep subband filtering for improved dereverberation," in *Proc. Interspeech*, 2023.

Chapter 4. Supervised Conditional Diffusion Models for Speech Dereverberation

4.1. Diffusion Models for Audio Restoration

[P5] J.-M. Lemercier, J. Richter, S. Welker, E. Moliner, V. Välimäki, and T. Gerkmann, "Diffusion models for audio restoration," *IEEE Signal Process. Magazine*, 2025.

[P6] 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 Lang. Process.*, vol. 31, pp. 2351–2364, 2023.

4.2. Analyzing Predictive Approaches versus Diffusion-based Generative Models for Speech Restoration

[P7] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing discriminative versus diffusion generative models for speech restoration tasks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023.

4.3. Combining Predictive Approaches and Diffusion-based Generative Models for Speech Enhancement and Dereverberation

[P8] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 31, pp. 2724–2737, 2023.

[P9] J.-M. Lemercier, J. Thiemann, R. Koning, and T. Gerkmann, "Wind noise reduction with a diffusion-based stochastic regeneration model," in *Proc. ITG Conf. Speech Communication*, 2023.

Chapter 5. Unsupervised Vocal Dereverberation and Room Acoustics Estimation with Diffusion Models

5.1. Unsupervised Diffusion Models for Informed Dereverberation

[P10] J.-M. Lemercier, S. Welker, and T. Gerkmann, "Diffusion posterior sampling for informed single-channel dereverberation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2023.

[P11] J.-M. Lemercier*, E. Moliner*, S. Welker, V. Välimäki, and T. Gerkmann, "Unsupervised blind joint dereverberation and room acoustics estimation with diffusion models," *arXiv:2408.07472. Submitted to IEEE/ACM Trans. Audio Lang. Speech Process.*, 2024.

5.2. Blind Dereverberation and Room Impulse Response Estimation with Unsupervised Diffusion Models

[P12] E. Moliner*, J.-M. Lemercier*, S. Welker, T. Gerkmann, and V. Välimäki, "BUDDy: Single-channel blind unsupervised dereverberation with diffusion models," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, 2024.

[P11] J.-M. Lemercier*, E. Moliner*, S. Welker, V. Välimäki, and T. Gerkmann, "Unsupervised blind joint dereverberation and room acoustics estimation with diffusion models," *arXiv:2408.07472. Submitted to IEEE/ACM Trans. Audio Lang. Speech Process.*, 2024.

[P13] É. Thuillier, J.-M. Lemercier, E. Moliner, T. Gerkmann, and V. Välimäki, "HRTF estimation using a score-based prior," *arXiv:2408.07472. Submitted to IEEE Int. Conf. Acoust. Speech Signal Process.*, 2024.

* Equal contribution

3

Model-based Speech Dereverberation

3.1 Lightweight Model-Based Dereverberation for Hearing Devices [P1]

Abstract

A two-stage lightweight online dereverberation algorithm for hearing devices is presented in this paper. The approach combines a multi-channel multi-frame linear filter with a single-channel single-frame post-filter. Both components rely on power spectral density (PSD) estimates provided by deep neural networks (DNNs). By deriving new metrics analyzing the dereverberation performance in various time ranges, we confirm that directly optimizing for a criterion at the output of the multi-channel linear filtering stage results in a more efficient dereverberation as compared to placing the criterion at the output of the DNN to optimize the PSD estimation. More concretely, we show that training this stage end-to-end helps further remove the reverberation in the range accessible to the filter, thus increasing the *early-to-moderate* reverberation ratio. We argue and demonstrate that it can then be well combined with a post-filtering stage to efficiently suppress the residual late reverberation, thereby increasing the *early-to-final* reverberation ratio. This proposed two stage procedure is shown to be both very effective in terms of dereverberation performance and computational demands, as compared to e.g. recent state-of-the-art DNN approaches. Furthermore, the proposed two-stage system can be adapted to the needs of different types of hearing-device users by controlling the amount of reduction of early reflections.

Reference

Jean-Marie Lemercier, Joachim Thiemann, Raphael Koning and Timo Gerkmann "A Neural Network-Supported Two-Stage Algorithm for Lightweight Dereverberation on Hearing Devices", *EURASIP Journal on Audio Speech and Music Processing*, vol. 18. 2023. DOI: 10.1186/s13636-023-00285-8

Copyright Notice

Authors' Contributions

Jean-Marie Lemercier is the first author of this publication. He implemented all algorithms, trained the neural networks used in the paper, conducted the experimental validation, and wrote the manuscript. Timo Gerkmann brought insights on the experimental validation, mathematical derivations, and also thoroughly reviewed the paper. Joachim Thiemann and Raphael Koning brought their feedback on all methods through discussions, they also helped with reviewing the manuscript.

RESEARCH

A Neural Network-supported Two-Stage Algorithm for Lightweight Dereverberation on Hearing Devices

Jean-Marie Lemercier[1*], Joachim Thiemann[2], Raphael Koning[2] and Timo Gerkmann[1]

Abstract

A two-stage lightweight online dereverberation algorithm for hearing devices is presented in this paper. The approach combines a multi-channel multi-frame linear filter with a single-channel single-frame post-filter. Both components rely on power spectral density (PSD) estimates provided by deep neural networks (DNNs). By deriving new metrics analyzing the dereverberation performance in various time ranges, we confirm that directly optimizing for a criterion at the output of the multi-channel linear filtering stage results in a more efficient dereverberation as compared to placing the criterion at the output of the DNN to optimize the PSD estimation. More concretely, we show that training this stage end-to-end helps further remove the reverberation in the range accessible to the filter, thus increasing the *early-to-moderate* reverberation ratio. We argue and demonstrate that it can then be well combined with a post-filtering stage to efficiently suppress the residual late reverberation, thereby increasing the *early-to-final* reverberation ratio. This proposed two stage procedure is shown to be both very effective in terms of dereverberation performance and computational demands, as compared to e.g. recent state-of-the-art DNN approaches. Furthermore, the proposed two-stage system can be adapted to the needs of different types of hearing-device users by controlling the amount of reduction of early reflections.

Keywords: dereverberation; neural network; end-to-end learning; hearing devices

1 Introduction

Communication and hearing devices require modules aiming at suppressing undesired parts of the signal to improve the speech quality and intelligibility. Reverberation is one of such distortions caused by room acoustics, and is characterized by multiple reflections on the room enclosures. Late reflections particularly degrade the speech signal and may result in a reduced intelligibility [1].

Traditional approaches were proposed for dereverberation such as spectral enhancement [2], beamforming [3], a combination of both [4], coherence weighting [5, 6], and linear-prediction based approaches such as the well-known weighted prediction error (WPE) algorithm [7, 8]. WPE computes an auto-regressive multi-channel filter in the short-time spectrum and applies it to a delayed group of reverberant speech frames. This approach is able to partially cancel late reverberation while inherently preserving parts of the early reflections, thus improving speech intelligibility for normal and hearing-supported listeners [9].

WPE and its extensions require the prior estimation of the anechoic speech PSD, which is modelled for instance through the speech periodogram [7] or a power-compressed periodogram corresponding to sparse priors [8], by an autoregressive process [10] or through non-negative matrix factorization [11]. A DNN was first introduced in [12] to model the anechoic PSD, thus avoiding the use of an iterative refinement.

Instead of providing parameters for linear prediction as in e.g. [12, 13], DNNs were also proposed for mapping-based dereverberation in the time-frequency magnitude domain [14], complex domain [15, 16] or in the time-domain [17].

As hearing devices operate in real-world scenarios in real-time, the proposed techniques for dereverberation should support low-latency online processing and adapt to changing room acoustics. Such online adaptive approaches were introduced, based on either Kalman filtering [18, 19] or on a recursive least squares (RLS)-adapted WPE, which is a special case

*Correspondence: jeanmarie.lemercier@uni-hamburg.de
[1]Signal Processing, Universität Hamburg, Hamburg, Germany
Full list of author information is available at the end of the article

of Kalman filtering [20]. Strategies for handling the case of speakers changing positions were introduced in [20, 19]. In the RLS-WPE framework, the PSD is either estimated by recursive smoothing of the reverberant signal [20] or by a DNN [21].

In the previously cited works, the DNN is trained towards PSD estimation, although this stage is only a front-end followed by RLS-WPE-based dereverberation algorithms. So-called "end-to-end techniques" aim to solve this mismatch by using a criterion placed at the output of the complete algorithm to train the DNN. End-to-end techniques using an automatic speech recognition (ASR) criterion were designed to refine the front-end DNN handling e.g. speech separation [22], denoising [23], or multiple tasks [24]. An end-to-end procedure using ASR as a training criterion was also introduced in [25] to optimize a DNN used for online dereverberation.

This journal paper is an extension of our prior work [26], where we proposed instead to use a criterion directly on the output signal rather than using ASR. We experimentally showed that it improved instrumentally predicted speech intelligibility and quality. The proposed criterion also enabled us to use different target signals and corresponding WPE parameters to make our approach adapt to the needs of different hearing-aid users categories: hearing aid (HA) users on the one hand benefiting from early reflections like normal listeners [9], and cochlear implant (CI) users on the other hand which do not benefit from early reflections [27].

We noticed in [26] that although the energy residing in the moderate reverberation range corresponding to the filter length was particularly suppressed when training the approach end-to-end, residual late reverberation could still be heard at the output. A further processing stage could be dedicated to removing this residual reverberation, as increasing the length of the linear filters results in rapidly increasing computational complexity. Hybrid approaches using such cascaded DNN-assisted stages have been proposed for dereverberation [28] or joint dereverberation, separation and denoising [29, 13, 24].

The extension to our work [26] consists in the three following contributions. First, we introduce metrics to measure the energies in various reverberation ranges in order to investigate the differences between the previously cited WPE-based approaches and our proposed method. Second, we propose to use a second DNN-supported stage based on single-frame non-linear magnitude filtering, and show that it significantly suppresses the residual late reverberation at the output of WPE. We show with the newly introduced metrics that this latter stage particularly benefits from strong dereverberation within the linear filter range obtained with the previous end-to-end WPE approach. Finally, we evaluate our approach and baselines on simulated reverberant data inspired by the WHAMR! dataset [30].

The rest of this paper is organized as follows. In Section 2, the online DNN-WPE dereverberation scheme is summarized. Section 3 presents the DNN-supported post-filter and describes the used end-to-end training procedure. In Section 4, we describe the experimental setup and introduce metrics in order to detail the dereverberation performance in various ranges. The results are presented and discussed in Section 5.

2 Signal model and DNN-supported WPE Dereverberation

2.1 Signal model

We use a subband-filtering approximation in the short-time Fourier transform (STFT) domain as in [7], and all computations except those involving neural networks are computed for each frequency band independently. Therefore, we omit the frequency index f when unnecessary and all vectors and matrices have an additional implicit frequency dimension of size F. The time frame index in the sequences of length T is denoted by t and is also dropped when not explicitly needed. We use lowercase normal font notation for signals having only time (and frequency) dimensions ($a_t \in \mathbb{C}$), lowercase bold font notation for vectors having one extra dimension ($\boldsymbol{a}_t \in \mathbb{C}^{d_1}$) and reserve uppercase bold font notation for matrices having two extra dimensions ($\boldsymbol{A}_t \in \mathbb{C}^{d_1 \times d_2}$).

The reverberant speech $\boldsymbol{x} \in \mathbb{C}^{D \times T}$ is obtained at the D-microphone array by convolution of the anechoic speech $s \in \mathbb{C}^T$ and the room impulse responses (RIRs) $\boldsymbol{h} \in \mathbb{C}^{D \times N}$:

$$\boldsymbol{x}_t = \sum_{\tau=0}^{N-1} \boldsymbol{h}_\tau s_{t-\tau} + \boldsymbol{u}_t = \boldsymbol{d}_t + \boldsymbol{e}_t + \boldsymbol{r}_t + \boldsymbol{u}_t, \qquad (1)$$

where $\boldsymbol{d}$ denotes the direct path, $\boldsymbol{e}$ the early reflections component, $\boldsymbol{r}$ the late reverberation and $\boldsymbol{u}$ an error term comprising modelling errors and background noise. The early reflections component $\boldsymbol{e}$ was shown to contribute to speech quality and intelligibility for normal and HA listeners [9] but not for CI users, particularly in highly-reverberant scenarios [27]. Therefore, we propose that the dereverberation objective is to retrieve $\boldsymbol{\nu} = \boldsymbol{d} + \boldsymbol{e}$ for HA listeners and $\boldsymbol{\nu} = \boldsymbol{d}$ for CI listeners.

2.2 WPE dereverberation

In relation to the subband reverberant model in (1), the WPE algorithm [7] uses an auto-regressive model

to approximate the late reverberation $\boldsymbol{r}$. Based on a zero-mean time-varying Gaussian model on the STFT anechoic speech s with time (and frequency) dependent PSD $\lambda^{(\mathrm{WPE})}$, a multi-channel filter $\boldsymbol{G} \in \mathbb{C}^{DK \times D}$ with K taps is estimated. This filter aims at representing the inverse of the late tail of the RIRs $\boldsymbol{h}$, such that the target $\boldsymbol{\nu}$ can be obtained through linear prediction with delay Δ. The prediction delay Δ is originally intended to avoid undesired short-time speech cancellations in [7], however this also leads to preserving parts of the early reflections. As such, we propose to set Δ larger for normal hearing and HA users who benefit from early reflections [9], but lower for CI users who suffer from early reflections [27]. By disregarding the error term $\boldsymbol{u}$ in (1) in noiseless scenarios, we obtain:

$$\boldsymbol{\nu}_t^{(\mathrm{WPE})} = \boldsymbol{x}_t - \boldsymbol{G}_t^H \mathcal{X}_{t-\Delta}, \tag{2}$$

where $\mathcal{X}_{t-\Delta} = \left[\boldsymbol{x}_{t-\Delta}^T, \ldots, \boldsymbol{x}_{t-\Delta-K+1}^T\right]^T \in \mathbb{C}^{DK}$.

In order to obtain an adaptive and real-time capable approach, RLS-WPE was proposed in [20], where the WPE filter $\boldsymbol{G}$ is recursively updated along time. RLS-WPE can be seen as a special case of Kalman filtering, in which the target covariance matrix is replaced by the scaled identity matrix $\lambda^{(\mathrm{WPE})}\boldsymbol{I}$, and the weight state error matrix is simply updated by dividing by the recursive factor α instead of following the usual Markov model [19]:

$$\boldsymbol{k}_t = \frac{(1-\alpha)\boldsymbol{R}_{t-1}^{-1}\mathcal{X}_{t-\Delta}}{\alpha\lambda_t^{(\mathrm{WPE})} + (1-\alpha)\mathcal{X}_{t-\Delta}^H \boldsymbol{R}_{t-1}^{-1}\mathcal{X}_{t-\Delta}}, \tag{3}$$

$$\boldsymbol{R}_t^{-1} = \frac{1}{\alpha}\boldsymbol{R}_{t-1}^{-1} - \frac{1}{\alpha}\boldsymbol{k}_t \mathcal{X}_{t-\Delta}^T \boldsymbol{R}_{t-1}^{-1}, \tag{4}$$

$$\boldsymbol{G}_t = \boldsymbol{G}_{t-1} + \boldsymbol{k}_t(\boldsymbol{x}_t - \boldsymbol{G}_{t-1}^H \mathcal{X}_{t-\Delta})^H. \tag{5}$$

$\boldsymbol{k} \in \mathbb{C}^{DK}$ is the Kalman gain, $\boldsymbol{R} \in \mathbb{C}^{DK \times DK}$ the covariance of the delayed reverberant signal buffer $\mathcal{X}_{t-\Delta}$ weighted by the PSD estimate $\lambda^{(\mathrm{WPE})}$, and α the forgetting factor.

In non-idealistic scenarios, the term $\boldsymbol{u}$ is not zero. Therefore, a regularization parameter $\epsilon > 0$ is added to the denominator of (3) which can be seen as a form of spectral flooring as used in traditional spectral enhancement schemes [31, 4, 6]. Although it is not *per se* a denoising solution and we still consider scenarios where noise is negligible in comparison to reverberation, adding this parameter helps increasing the robustness of WPE to noise, numerical instabilities and modelling errors. On the other hand, setting ϵ to a high value will excessively attenuate the relative variations of the Kalman denominator, which mitigates the benefits of variance-normalization as explained in [32]. A value of $\epsilon^* = 0.001$ was picked based on the performance of the WPE algorithm using oracle PSD.

2.3 DNN-based PSD estimation

The anechoic speech PSD estimate $\lambda^{(\mathrm{WPE})}$ is obtained at each time step, either by recursive smoothing of the reverberant periodogram [20] or with help of a DNN [21]. A block diagram of the DNN-WPE algorithm as proposed in [21] is given in Figure 1, as the first stage up to $\boldsymbol{\nu}^{(\mathrm{WPE})}$. In this approach, the input to the neural network is the magnitude of the reference channel $|x_0|$, taken here to be the first channel. We did not observe changes in the results by changing the reference channel or computing an average of the channels to obtain the DNN input, likely because the signal model itself considers a channel-agnostic PSD. The magnitude frame is then fed to a recurrent neural network $\mathrm{MaskNet}_{\mathrm{WPE}}$, which outputs a real-valued mask $\mathcal{M}^{(\mathrm{WPE})}$. The PSD estimate is obtained by time-frequency masking:

$$\lambda_{t,f}^{(\mathrm{WPE})} = (\mathcal{M}_{t,f}^{(\mathrm{WPE})} \odot |x_{0,t,f}|)^2, \tag{6}$$

where $\odot$ represents the Hadamard element product.

In [12, 21], the DNN is optimized with a mean-squared error (MSE) criterion on the masked output. In contrast, we proposed to use the L^1 loss:

$$\mathcal{L}_{\text{DNN-WPE}} = \sum_{t,f} \left| \mathcal{M}_{t,f}^{(\mathrm{WPE})} \odot |x_{0,t,f}| - |\nu_{0,t,f}| \right|. \tag{7}$$

This loss function indeed led to better results in our experiments [26]. This can be explained by the fact that the L^1 loss puts more weight on low-energy bins than high-energy bins in comparison to the MSE loss as it is more concave, which is a good fit for dereverberation.

2.4 End-to-End Training Procedure

2.4.1 End-to-end criterion and objectives

We argue that the mismatch between the DNN-optimization criterion (7) and the dereverberation task may limit the overall performance. However, using ASR as an end-to-end training criterion, as is done in [25], may not necessarily the best choice in order to optimize a dereverberation algorithm for hearing-aid users. The first reason is that the resulting scheme could not be adapted to specific user categories, although these benefit from different speech cues. Namely, HA listeners are shown to benefit from early reflections [9] where CI listeners do not significantly benefit from those, in particular in highly reverberant scenarios where early reflections degrade intelligibility [27]. The second reason is that by nature, the

dereverberation scheme will provide the best representation possible for ASR, which may be not the optimal representation in terms of quality and intelligibility for a human listener.

We therefore proposed an end-to-end training procedure where the optimization criterion is placed in the time-frequency domain at the output of the DNN-WPE algorithm, thus including the back-end WPE into DNN optimization:

$$\mathcal{L}_{\text{E2E-WPE}} = \sum_{t,f} \big| \, |\boldsymbol{\nu}_{t,f}^{(\text{WPE})}| - |\boldsymbol{\nu}_{t,f}| \, \big|. \tag{8}$$

2.4.2 End-to-end training procedure

An important practical aspect of this study focuses on handling the initialization period of the RLS-WPE algorithm. During this interval, the filter $\boldsymbol{G}$ has not yet converged to a stable value, reducing dereverberation performance. Therefore, rather than relying on a hypothetical shortening of this period through implicit PSD optimization [25], we choose to exclude this initialization period from training. The DNN is thus optimized so that the algorithm works best in its stable regime. To do so, we first craft long reverberant utterances that we cut in segments of L_i frames, where L_i is the worst case initialization time plus some margin. We then design the training procedure so that the first segment is used only to initialize the WPE statistics $\boldsymbol{G}$ and $\boldsymbol{R^{-1}}$ and the DNN hidden states $h(\text{MaskNet}_{\text{WPE}})$. This enables to train the DNN weights on the next segments, during the stable regime. The data generation procedure is detailed again in subsection 4.

We showed in [26] that the best performance was obtained with the **E2Ep-WPE** approach, where the network $\text{MaskNet}_{\text{WPE}}$ is first pre-trained with (7) and fine-tuned with (8). If $\text{MaskNet}_{\text{WPE}}$ is only pre-trained, the algorithm is named **DNN-WPE**, and corresponds to [21] with a different training loss function.

The proposed end-to-end training procedure is summarized in Algorithm 1.

Algorithm	$\lambda^{(\text{WPE})}$	$\lambda^{(\nu,\text{PF})}, \lambda^{(\tilde{r},\text{PF})}$
RLS-WPE [20]	Reverberant	✗
O-PSD-WPE	Oracle	✗
DNN-PF	✗	$\mathcal{L}_{\text{DNN-PF}}$
DNN-WPE	$\mathcal{L}_{\text{DNN-WPE}}$	✗
E2Ep-WPE	$\mathcal{L}_{\text{DNN-WPE}} \rightarrow \mathcal{L}_{\text{E2E-WPE}}$	✗
DNN-WPE+DNN-PF	$\mathcal{L}_{\text{DNN-WPE}}$	$\mathcal{L}_{\text{DNN-PF}}$
E2Ep-WPE+DNN-PF	$\mathcal{L}_{\text{DNN-WPE}} \rightarrow \mathcal{L}_{\text{E2E-WPE}}$	$\mathcal{L}_{\text{DNN-PF}}$

Table 1 List of acronyms for strategies estimating the PSD used in the linear filtering and non-linear post-filtering stages.

3 Residual Reverberation Suppression

3.1 Signal model

As shown in Section 5 below, training the DNN-supported WPE stage in an end-to-end fashion helps suppressing large part of the reverberant signal immediately following the target range, that is, up to L_m, which we refer to as the *moderate reverberation* range.

We thus refine the reverberant signal model as (1):

$$\boldsymbol{x}_t = \boldsymbol{\nu}_t + \boldsymbol{m}_t + \boldsymbol{\phi}_t + \boldsymbol{u}_t, \tag{9}$$

where the undesired reverberant signal in (1) (corresponding to $\boldsymbol{r}$ and $\boldsymbol{e}+\boldsymbol{r}$ in the HA and CI case respectively) is split in the *moderate* reverberant signal $\boldsymbol{m}$ and the *final* reverberant signal $\boldsymbol{\phi}$, defined as:

$$\boldsymbol{m}_t = \sum_{\tau=\Delta}^{\Delta+L_m-1} \boldsymbol{h}_\tau s_{t-\tau}, \tag{10}$$

$$\boldsymbol{\phi}_t = \sum_{\tau=\Delta+L_m}^{N} \boldsymbol{h}_\tau s_{t-\tau}. \tag{11}$$

The resulting WPE estimate thus contains the target $\boldsymbol{\nu}$, a target estimation error $\tilde{\boldsymbol{\nu}}$, a residue $\tilde{\boldsymbol{m}}$ from this moderate reverberation and a residue stemming from the final reverberation $\tilde{\boldsymbol{\phi}}$ (again disregarding the error term $\boldsymbol{u}$ in noiseless scenarios):

$$\begin{aligned} \boldsymbol{\nu}_t^{(\text{WPE})} &= \boldsymbol{x}_t - \boldsymbol{G}_t^H \mathcal{X}_{t-\Delta} \\ &= \boldsymbol{\nu}_t + \underbrace{\tilde{\boldsymbol{\nu}}_t + \tilde{\boldsymbol{m}}_t + \tilde{\boldsymbol{\phi}}_t}_{\tilde{\boldsymbol{r}}_t}. \end{aligned} \tag{12}$$

The target estimation error $\tilde{\boldsymbol{\nu}}$ is the target component which was degraded by the algorithm. As described in [32] for the original WPE algorithm, parts of the early reflections may be destroyed because of the inner short-time speech correlations. Under some mild assumptions, the direct path is however fully preserved if the prediction delay Δ is sufficiently large (i.e. larger than the inner speech correlation time). The target estimation error is therefore likely to be larger when using WPE-based algorithms in the HA scenario—containing more early reflections—than in the CI scenario.

3.2 Postfiltering scheme

We aim at suppressing the two residues $\tilde{\boldsymbol{m}}$ and, more particularly, $\tilde{\boldsymbol{\phi}}$. Indeed, $\tilde{\boldsymbol{\phi}}$ is generally of higher magnitude than $\tilde{\boldsymbol{m}}$, as we will show in the experiments that a large amount of moderate reverberation can be cancelled by efficient WPE-based dereverberation. Additionally, $\tilde{\boldsymbol{\phi}}$ is the more perceptually disturbing of the two residues for the following reasons.

On the one hand, $\tilde{\boldsymbol{\phi}}$ can be considered as speech-like noise which is very poorly correlated to the target signal in comparison to $\tilde{\boldsymbol{m}}$. On the other hand, as WPE cancels most of the so-called moderate reverberation,

there is no preceding energy anymore to mask the late reverberation. The final reverberation residue is then clearly audible.

We thus add a post-filtering enhancement stage after the linear WPE filtering stage, which consists of a single-channel Wiener filter, the phase being left unchanged. This Wiener filter uses estimates of the target PSD $\lambda^{(\nu,\mathrm{PF})}$ and interference PSD $\lambda^{(\tilde{r},\mathrm{PF})}$, which can be obtained with classical techniques as decision-directed signal-to-noise ratio (SNR) estimation [33], cepstral smoothing [34, 6], or from a neural network [21, 35].

The resulting estimate is then given for each channel d separately by the celebrated Wiener filter, using the WPE output:

$$\boldsymbol{\nu}_{d,t}^{(\mathrm{PF})} = \frac{\lambda_{d,t}^{(\nu,\mathrm{PF})}}{\lambda_{d,t}^{(\nu,\mathrm{PF})} + \lambda_{d,t}^{(\tilde{r},\mathrm{PF})}} \boldsymbol{\nu}_{d,t}^{(\mathrm{WPE})} \tag{13}$$

3.3 DNN-based PSD Estimation

We use a DNN-based masking approach to obtain the target and residual reverberation PSDs, similar to what is used to estimate the target speech PSD for WPE filtering (see (6)). At each time step, a frame of the WPE output's magnitude taken from the reference channel $|\nu_0^{(\mathrm{WPE})}|$ is fed to a recurrent neural network $\mathrm{MaskNet}_{\mathrm{PF}}$, which outputs both a target and interference mask. The PSD estimate $\lambda^{(\eta)}$ is then obtained for each channel d through time-frequency masking for each signal $\boldsymbol{\eta} \in \{\boldsymbol{\nu}, \tilde{\boldsymbol{r}}\}$:

$$\lambda_{d,t,f}^{(\eta)} = (\mathcal{M}_{t,f}^{(\eta)} \odot |\nu_{d,t,f}^{(\mathrm{WPE})}|))^2. \tag{14}$$

We apply the same reference-channel mask for all channels using only one instance of the DNN, which saves some computational power and enables us to leave the interaural level differences unchanged. Also, the interaural phase differences are well estimated by WPE linear filtering and are not modified by the post-filtering scheme (see (13)). Therefore the target binaural cues are well preserved, which is important for hearing devices.

A block diagram of the complete two-stage algorithm is provided in Figure 1.

3.4 Training Procedure

We trained the post-filter DNN $\mathrm{MaskNet}_{\mathrm{PF}}$ with a similar mask-based objective as $\mathrm{MaskNet}_{\mathrm{WPE}}$:

$$\mathcal{L}_{\text{DNN-PF}} = \sum_{t,f} \big| \, \mathcal{M}_{t,f}^{(\nu)} \odot |\nu_{0,t,f}^{(\mathrm{WPE})}| - |\nu_{0,t,f}| \, \big| + \sum_{t,f} \big| \, \mathcal{M}_{t,f}^{(\tilde{r})} \odot |\nu_{0,t,f}^{(\mathrm{WPE})}| - |\tilde{r}_{0,t,f}| \, \big|, \tag{15}$$

where $\tilde{r}_0$ is the undesired signal defined in (12) taken at the reference channel. We report results for two approaches. First is **DNN-WPE+DNN-PF**, where the network $\mathrm{MaskNet}_{\mathrm{WPE}}$ is pre-trained with (7), then frozen for the pre-training of $\mathrm{MaskNet}_{\mathrm{PF}}$ with (15). Second is **E2Ep-WPE+DNN-PF**, where the network $\mathrm{MaskNet}_{\mathrm{WPE}}$ is pre-trained with (7) and fine-tuned with (8), then frozen for the pre-training of $\mathrm{MaskNet}_{\mathrm{PF}}$ with (15).

A table making the present algorithms correspond to their characteristics and acronyms is given in Table 1.

Algorithm 1 End-to-End Training Procedure

Extract STFT of given sequence
Segment sequence in N segments of size L_i
for $n \in \{0 \dots N-1\}$ **do**
  if $n = 0$ **then** ▷ Initialization period
    Initialize LSTM hidden state:
    $h(\mathrm{MaskNet}_{\mathrm{WPE}})_0^{(0)} = 0$
    Initialize WPE statistics:
    $\boldsymbol{R^{-1}}_0^{(0)} = \boldsymbol{I}$; $\boldsymbol{G}_0^{(0)} = \boldsymbol{0}$
    for $t \in \{0 \dots L_i - 1\}$ **do**
      Compute $\boldsymbol{\nu}_t^{(\mathrm{WPE})}$
  if $n > 0$ **then** ▷ After initialization
    Forward LSTM hidden state:
    $h(\mathrm{MaskNet}_{\mathrm{WPE}})_0^{(n)} = h(\mathrm{MaskNet}_{\mathrm{WPE}})_{L_i-1}^{(n-1)}$
    Forward WPE statistics:
    $\boldsymbol{R^{-1}}_0^{(n)} = \boldsymbol{R^{-1}}_{L_i-1}^{(n-1)}$; $\boldsymbol{G}_0^{(n)} = \boldsymbol{G}_{L_i-1}^{(n-1)}$
    for $t \in \{0 \dots L_i - 1\}$ **do**
      Compute $\boldsymbol{\nu}_t^{(\mathrm{WPE})}$
    Backpropagate loss (8) through time on n

4 Experimental Setup

4.1 Dataset generation

We use clean speech material from the WS0 dataset [36], using the usual split of 101, 10 and 8 speakers for training, validation and testing respectively. For each split independently, we concatenate utterances belonging to the same speaker, and construct sequences of approximately 20 seconds. The initialization time of WPE can go up to to 2 seconds in the worst case when using a forgetting factor of $\alpha = 0.99$. For end-to-end training, we do not want to learn during that period (cf Section 2.4). Therefore we cut these long sequences in segments of $L_i = 4$ seconds and use the first segment only for initialization, thus not backpropagating the loss on it (cf Algorithm 1). We choose L_i to fill both requirements of (i) being larger than the worst case

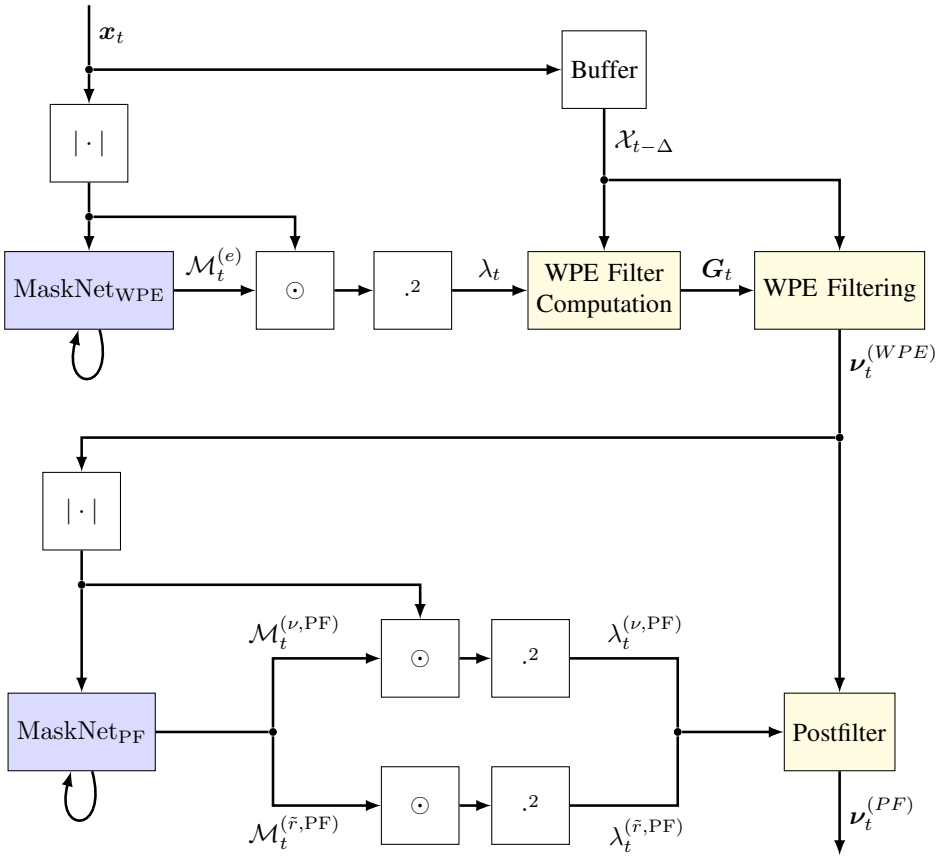


Figure 1 DNN-supported two-stage dereverberation. Blue blocks refer to trainable neural network layers. Yellow blocks represents adaptive statistical signal processing.

initialization time of WPE and (ii) providing a sufficient receptive field for training with LSTMs. Since the first segment is never used for optimization, permutations of the original utterances are used to create several versions of each sequence, so that we still use all speech data available for training the DNNs.

These sequences are convolved with 2-channel RIRs generated with the RAZR engine [37] and randomly picked. Each RIR is generated by uniformly sampling room acoustics parameters as in [30] and a T_{60} reverberation time between 0.4 and 1.0 seconds. Head-Related Transfer Function based auralization is performed in the RAZR engine, using a KEMAR dummy head response from the MMHR-HRTF database [38].

As specified earlier, the target data for the HA case should represent the direct path and the early reflections as normal hearing and hearing-aided listeners benefit from early reflections [9]. Therefore, we convolve the dry utterance with the beginning of the RIR, up to a separation time often found in the dereverberation literature [1, 39, 9]. We empirically set the separation time to 40 ms instead of the usual 50 ms, as we obtained better instrumental results when comparing the resulting target data to WPE estimates using the oracle PSD.

In the CI scenario, the target data data should theoretically contain the direct path only [27]. However, directly estimating the direct path from reverberant speech often provides poor instrumental results given the low input SNR. Note also that the first WPE stage uses a prediction delay Δ supposed to protect the inner speech correlations, whose range is usually estimated to ~ 10 ms. The minimal Δ that fills this requirement is $\Delta = 2$ STFT frames with the hyperparameters described below, that is, 16 ms. Therefore, we propose to match the target data with the best possible WPE estimate, by convolving the dry utterance with the first 16 ms of the RIR. This also contributes to decreasing the difficulty of the estimation task, which helps obtain reasonable estimates with the proposed algorithm. We

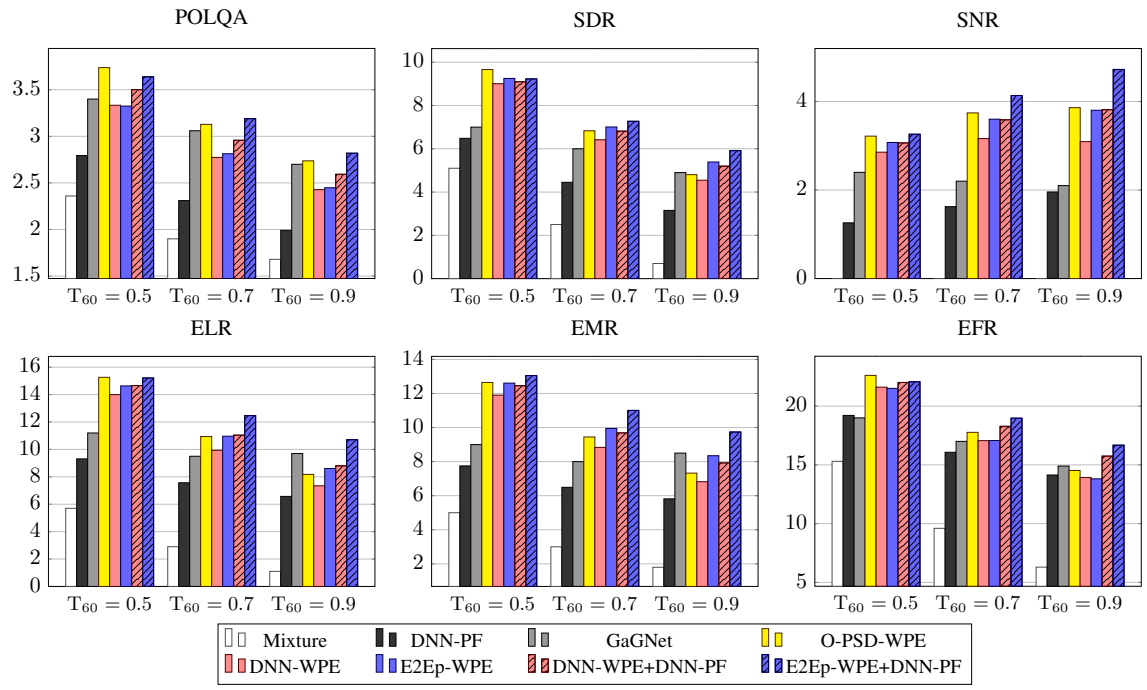


Figure 2 Scores on unprocessed and processed signals for hearing-aided scenario. All metrics except POLQA are in dB. T_{60} times indicated in s. [$\boldsymbol{\nu} = \boldsymbol{d} + \boldsymbol{e}$; $\Delta = 5$]

further noticed that with this setting, very few early reflections could be heard in the target.

The original mean input direct-to-reverberant ratio (DRR) between the dry signal and reverberant mixture is −6.0dB and the mean microphone-to-speaker distance used was estimated to 4.2m. The resulting mean input signal-to-noise ratio (SNR) between the generated target and the reverberant mixture is 0.9dB for the HA scenario, and −1.4dB for the CI scenario.

Finally, independent and identically distributed Gaussian noise is added to each channel with an input SNR uniformly sampled in $[15, 25]$ dB to simulate sensor noise. Ultimately, the training, validation and testing sets contain around 55, 16 and 3 hours of speech sampled at 16 kHz.

4.2 Hyperparameter settings

The STFT uses a square-rooted Hann window of 32 ms and a 75 % overlap. For training, segments of $L_i = 4$ s are constructed from each sequence (see Section 4.1). All approaches are trained using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 128. Training is stopped if a maximum of 500 epochs is reached or if early stopping is detected, in case the validation loss has not decreased in 20 consecutive epochs.

The WPE filter length is set to $K = 10$ STFT frames (i.e. 80 ms), the number of channels to $D = 2$, the WPE adaptation factor to $\alpha = 0.99$ and the delays to $\Delta_{\text{HA}} = 5$ frames (i.e. 40 ms) for the HA scenario and $\Delta_{\text{CI}} = 2$ (i.e. 16 ms) frames for the CI scenario. The delay values are picked to match the amount of early reflections contained in the respective target, and they experimentally provide optimal evaluation metrics when comparing the corresponding target to the output of WPE when using the oracle PSD (see Section 4.1).

The DNN used in [21] is composed of a single long-short term memory (LSTM) layer with 512 units followed by two linear layers with rectified linear activations (ReLU), and a linear output layer with sigmoid activation. We remove the two ReLU-activated layers in our experiments, which did not significantly degrade the dereverberation performance, while reducing the number of trainable parameters by 75 %, therefore ending with 1.6M parameters. We use the same architecture for $\text{MaskNet}_{\text{WPE}}$ and $\text{MaskNet}_{\text{PF}}$. We choose to use LSTMs rather than recent convolutional network- or transformer-based architectures to develop a frugal algorithm for hearing devices with limited computing resources. Indeed, LSTMs require much fewer operations per second than the mentioned alternatives, given that they process only one input frame and perform sequence-modelling using their internal memory state.

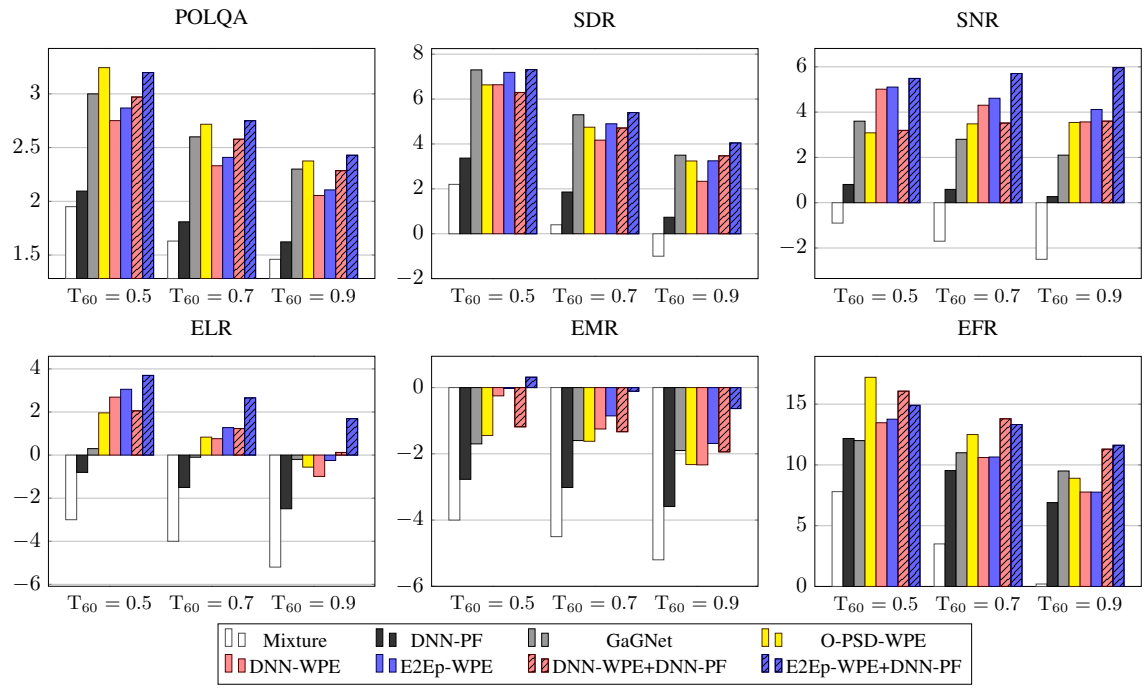


Figure 3 Scores on unprocessed and processed signals for cochlear-implanted scenario. All metrics except POLQA are in dB. T_{60} times indicated in s. [$\boldsymbol{\nu} = \boldsymbol{d}$; $\Delta = 2$]

4.3 Evaluation metrics

We evaluate all approaches on the described test sets corresponding to the HA and CI scenarios.

Following the definition of the early-to-late reverberation ratio (ELR) [40, 10], we introduce two new instrumental measures: the *early-to-moderate reverberation ratio* (EMR) and *early-to-final reverberation ratio* (EFR). Estimated RIR coefficients $\{\hat{H}\}_{d,\tau,f}$ of order $0 \leq \tau \leq P-1$ are computed for each channel d and frequency bin f separately, in order to minimize a minimum mean square error regression objective in the time-frequency domain between a reverberant utterance Y and the corresponding dry utterance S filtered by H [13]:

$$\{\hat{H}_{d,\tau,f}\}_\tau = \arg\min_H \sum_{t=0}^{T-1} ||Y_{d,t,f} - \sum_{\tau=0}^{P-1} H_{d,\tau,f} S_{t-\tau-\delta^*,f}||_2^2, \quad (16)$$

with δ^* being the oracle propagation delay obtained by looking for the direct path in the true RIR. This delay is used so as not to try and estimate RIR coefficients preceding the propagation delay which are supposed to be zero, therefore reducing the estimation error. The estimation error is further reduced by choosing the order P to match the T_{30} of the true RIR rather than the T_{60}, as the estimation error floor was found to be close to -30dB.

The channel-wise RIRs are then stacked and the target, moderate and final reverberation components are estimated as:

$$\widehat{\boldsymbol{\nu}}_{t,f} = \sum_{\tau=0}^{\tilde{\Delta}-1} \hat{H}_{\tau,f} S_{t-\tau-\delta^*,f}, \quad (17)$$

$$\widehat{\boldsymbol{m}}_{t,f} = \sum_{\tau=\tilde{\Delta}}^{\tilde{\Delta}+L_m-1} \hat{H}_{\tau,f} S_{t-\tau-\delta^*,f}, \quad (18)$$

$$\widehat{\boldsymbol{\phi}}_{t,f} = \sum_{\tau=\tilde{\Delta}+L_m}^{P-1} \hat{H}_{\tau,f} S_{t-\tau-\delta^*,f}. \quad (19)$$

We set $\tilde{\Delta} = 5$ (i.e. 40ms) in the hearing-aided case and $\tilde{\Delta} = 2$ (i.e. 16ms) in the cochlear-implanted scenario as explained in the target specifications in the section above. We set the moderate range length to $L_m = K = 10$ (i.e. 80ms).

The ELR, EMR and EFR are then defined as:

$$\text{ELR} = 10\log_{10}\left(||\hat{\boldsymbol{\nu}}||^2 / ||\hat{\boldsymbol{m}} + \hat{\boldsymbol{\phi}}||^2\right), \quad (20)$$

$$\text{EMR} = 10\log_{10}\left(||\hat{\boldsymbol{\nu}}||^2 / ||\hat{\boldsymbol{m}}||^2\right), \quad (21)$$

$$\text{EFR} = 10\log_{10}\left(||\hat{\boldsymbol{\nu}}||^2 / ||\hat{\boldsymbol{\phi}}||^2\right). \quad (22)$$

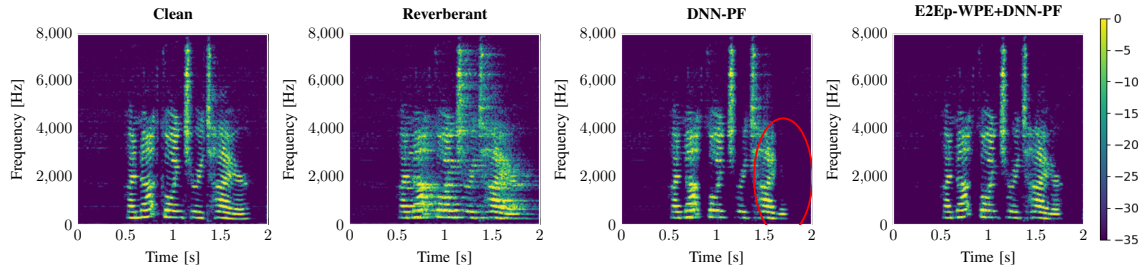


Figure 4 Log-energy spectrograms of clean, reverberant and processed utterances. $T_{60} = 0.68$s. HA scenario [$\boldsymbol{\nu} = \boldsymbol{d} + \boldsymbol{e}$; $\Delta = \tilde{\Delta} = 5$] Heavy speech distortions can be observed in the DNN-PF output, as highlighted in the red ellipse.

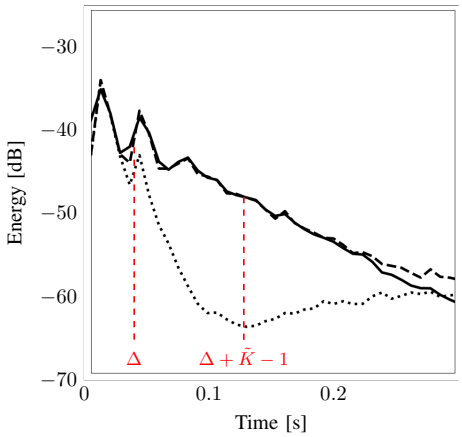


Figure 5 Comparison of the true RIR (full line) vs. the Estimated RIR (dashed line). Estimated linearized transfer function of the system which applies O-PSD-WPE on reverberant speech is shown as a dotted line. We observe strong dereverberation in the given filter range $[\Delta, \Delta + \tilde{K} - 1]$, and shortly afterwards because of recursive averaging. Only the T_{30} range is displayed as it is the valid estimation range for the estimated RIR. $T_{60} = 0.8s$. HA scenario [$\boldsymbol{\nu} + \boldsymbol{d} + \boldsymbol{e}$; $\Delta = \tilde{\Delta} = 5$]

We complete the evaluation benchmark with Perceptual Objective Listening Quality Analysis (POLQA)[1], signal-to-distortion ratio (SDR) and signal-to-noise ratio [41].

5 Experimental Results and Discussion

5.1 Compared algorithms

We apply the different strategies mentioned in sections 2 and 3 and compare their results in figures 2 and 3 for the HA and CI scenarios of our simulated dataset respectively.

[1] Wideband MOS score, following standard ITU-T P.863. The authors would like to thank Rohde & Schwarz SwissQual AG for their support with POLQA.

Spectrograms are also plotted in Figure 4. We add to the already proposed approaches (mentioned in bold):

- **O-PSD-WPE**: RLS-WPE using the oracle target PSD.
- **DNN-PF**: The output of the network $\text{MaskNet}_{\text{WPE}}$ is directly used for single-channel Wiener nonlinear filtering, eluding the WPE linear filter step.
- **GaGNet** [42]: A recent CNN-based network for hybrid magnitude and complex domain enhancement. GaGNet is the successor of [43] which was ranked first in the real-time enhancement track of the DNS-2021 challenge [44]. We used the open source available implementation[2], but adapted the number of frequency bins to be 257 as in our implementation.

Some listening examples and spectrograms are available on our dedicated webpage[3]. We also include there a video recording of our proposed E2Ep-WPE+DNN-PF (HA) algorithm performing in real time in both static and moving speaker scenarios. The algorithm performs with a total latency of 40ms determined by the 32ms *algorithmic latency* due to the STFT synthesis window length and the 8ms *processing time* which is contained within a STFT hop. We show that for reasonable speaker movements, the algorithm yields high performance also in the dynamic setting.

5.2 Moderate Reverberation Suppression

We first validate the method used for deriving the ELR, EMR and EFR metrics, described in 4.3. We plot the log-energies of the true RIR, the RIR estimated with (16) and the transfer function of the concatenation of the room with the O-PSD-WPE algorithm on Fig. 5. We observe that in the chosen T_{30} range, the true and estimated RIRs match almost perfectly, showing the validity of this MMSE-based estimation for linear transfer function estimation in this range.

[2] https://github.com/Andong-Li-speech/GaGNet
[3] https://uhh.de/inf-sp-twostagederev

We also observe a strong derverberation performance of the O-PSD-WPE algorithm in the filter range as well as shortly after this range, which is the effect of recursive averaging.

The ELR metric in figures 2 and 3 indicates a superior dereverberation performance of E2Ep-WPE in comparison to DNN-WPE, i.e. when the DNN $\text{MaskNet}_{\text{WPE}}$ is fine-tuned end-to-end. The high EMR difference indicates that the moderate reverberation in the range $[\tilde{\Delta}, \tilde{\Delta} + L_m - 1]$ is particularly well suppressed. As already mentioned in [26], this stems from the better dereverberation performance in the range which is available to the WPE linear filter, through end-to-end optimization of the neural network $\text{MaskNet}_{\text{WPE}}$.

5.3 Residual Reverberation Suppression

As displayed in figures 2 and 3, using a DNN-assisted post-filtering stage highly improves the dereverberation performance on the basis of WPE linear filtering, and yields much superior POLQA scores. The high EFR improvement indicates that post-filtering mostly focuses on removing the final reverberation, i.e. after the range accessible to WPE filtering. In particular, the E2Ep-WPE+DNN-PF approach which uses a pretrained network for post-filtering on top of end-to-end trained WPE filtering, outperforms all other approaches on all metrics. In comparison, using only the post-filter without WPE filtering introduces a lot of speech distortion, as shown in Figure 4. Similarly, the DNN-WPE+DNN-PF performance indicates that using the post-filtering stage on the output of the DNN-WPE algorithm—without fine-tuning $\text{MaskNet}_{\text{WPE}}$ with our end-to-end procedure—yields poorer results (final POLQA is 0.2 lower and SNR is 1dB lower than E2Ep-WPE+DNN-PF). This shows that removing the moderate reverberation with WPE linear filtering is an essential step before using a speech enhancement scheme like our post-filter. Since E2Ep-WPE efficiently removes the moderate reverberation, as measured by EMR, it provides a particularly good ground for enhancement-like post-filtering, since only the reverberation tail remains, and provides the best EFR and POLQA performance.

5.4 Reverberation Times

For a given scenario, the dereverberation task becomes increasingly difficult as the T_{60} time grows longer. We observe for example that using the oracle PSD for WPE performs well only for low T_{60} reverberation times because of the limited filter length, and the performance gap between this approach and the proposed two-stage approach increases with the T_{60} reverberation time.

Furthermore, we notice an increasing gap in SNR and EFR between DNN-WPE+DNN-PF and E2Ep-WPE+DNN-PF as the T_{60} grows larger, which seems to indicate that our best performing approach E2Ep-WPE+DNN-PF is more robust to challenging reverberation conditions.

5.5 Hearing Device Users Categories Specialization

Similar trends in performances are observed for the hearing-aided and cochlear-implanted scenarios.

Dereverberation is a more complicated task in the CI scenario as compared to the HA scenario, as the input ELR and SDR scores are lower. Yet, the POLQA and SDR score improvements stay relatively consistent across both scenarios, highlighting the robustness of our approach. However, the EMR improvements seem larger in the HA scenario than in the CI scenario. Indeed, it is more arduous in the latter scenario to remove the beginning of what is considered to be the reverberant tail, as it includes parts of the early reflections, which are complicated to attenuate without degrading the direct path. This also accounts for the smaller EMR improvement of E2Ep-WPE over DNN-WPE, as compared to the HA scenario. Furthermore, the SNR improvements are larger in the CI scenario than in the HA scenario, especially those brought by the proposed E2Ep-WPE+DNN-PF approach, which shows that the post-filtering stage is in this case able to remove a lot of the residual reverberation.

5.6 Computational requirements

We estimate the number of MAC operations per second of the models using the `python-papi` Python package which provides CPU counters for single- and double-point precision operations. We end up with an estimate of 0.13 GMAC$\cdot$s^{-1} for our proposed E2Ep-WPE+DNN-PF algorithm running at 16 kHz. With the same estimation method, the implemented GaGNet uses 0.81 GMAC$\cdot$s^{-1} . Also with regard to memory, our method has a lower budget as GaGNet has 11.8M trainable parameters while our approach has 3.2M parameters.

Our method therefore outperforms GaGNet on the proposed dataset with a significantly smaller computational load, without special fine-tuning of the hyperparameters nor optimization of the architectures used.

6 Conclusions

We have proposed a lightweight two-stage DNN-assisted algorithm for frame-online adaptive multi-channel dereverberation on hearing devices. The first stage consists of multi-frame, multi-channel linear filtering with help of a DNN estimating the target speech PSD, optimized end-to-end. This first stage was shown

to focus on accurately removing moderate reverberation up to the given filter range, in our case, 120 ms. The second stage performs channel-wise, single-frame non-linear spectral enhancement with help of a DNN estimating the target and interference PSDs. This second stage is able to efficiently remove residual late reverberation left off by the first stage.

Our model-based approach allows to tailor the two-stage algorithm toward different classes of hearing-impaired listeners, namely hearing-impaired listeners benefiting from early reflections on the one hand, and cochlear-implanted users on the other hand benefiting from the direct path only.

Instrumental metrics like the early-to-late reverberation ratio and its variants confirm the listening-based experiments showing the complementary aspect of the two proposed stages.

The proposed approach outperforms a state-of-the-art DNN-based enhancement scheme on the proposed dataset, using a significantly smaller time and memory footprint.

Funding
This work has been funded by the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380. The authors are responsible for the content of this paper.

Author details
[1]Signal Processing, Universität Hamburg, Hamburg, Germany. [2]Advanced Bionics, Hannover, Germany.

References

1. Naylor, P., Gaubitch, N.: Speech dereverberation. Noise Control Engineering Journal **59** (2011)
2. Habets, E.: Single- and multi-microphone speech dereverberation using spectral enhancement. PhD thesis (2007)
3. Kuklasiński, A., Doclo, S., Gerkmann, T., Holdt Jensen, S., Jensen, J.: Multi-channel PSD estimators for speech dereverberation - a theoretical and experimental comparison. In: IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP) (2015)
4. Cauchi, B., Kodrasi, I., Rehr, R., Gerlach, S., Jukic, A., Gerkmann, T., Doclo, S., Goetze, S.: Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech. EURASIP Journal on Advances in Signal Proc. **2015** (2015)
5. Schwarz, A., Kellermann, W.: Coherent-to-diffuse power ratio estimation for dereverberation. IEEE/ACM Trans. Audio, Speech, Language Proc. **23**(6) (2015)
6. Gerkmann, T.: Cepstral weighting for speech dereverberation without musical noise. In: Proc. Euro. Signal Proc. Conf. (EUSIPCO) (2011)
7. Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B.: Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation. In: IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP) (2008)
8. Jukić, A., van Waterschoot, T., Gerkmann, T., Doclo, S.: Multi-channel linear prediction-based speech dereverberation with sparse priors. IEEE/ACM Trans. Audio, Speech, Language Proc. **23**(9) (2015)
9. Bradley, J.S., Sato, H., Picard, M.: On the importance of early reflections for speech in rooms. The Journal of the Acoustical Society of America **113**(6) (2003)
10. Yoshioka, T., Nakatani, T., Miyoshi, M., Okuno, H.G.: Blind separation and dereverberation of speech mixtures by joint optimization. IEEE Trans. Audio, Speech, Language Proc. **19**(1) (2011)
11. Kagami, H., Kameoka, H., Yukawa, M.: Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization. In: IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP) (2018)
12. Kinoshita, K., Delcroix, M., Kwon, H., Mori, T., Nakatani, T.: Neural network-based spectrum estimation for online WPE dereverberation. In: ISCA Interspeech (2017)
13. Wang, Z.-Q., Wichern, G., Roux, J.L.: Convolutive prediction for monaural speech dereverberation and noisy-reverberant speaker separation. IEEE/ACM Trans. Audio, Speech, Language Proc. **29** (2021)
14. Han, K., Wang, Y., Wang, D., Woods, W.S., Merks, I., Zhang, T.: Learning spectral mapping for speech dereverberation and denoising. IEEE/ACM Trans. Audio, Speech, Language Proc. **23**(6) (2015)
15. Williamson, D.S., Wang, D.: Time-frequency masking in the complex domain for speech dereverberation and denoising. IEEE/ACM Trans. Audio, Speech, Language Proc. **25**(7) (2017)
16. Li, A., Zheng, C., Zhang, L., Li, X.: Glance and gaze: A collaborative learning framework for single-channel speech enhancement. Applied Acoustics **187** (2022)
17. Luo, Y., Mesgarani, N.: Real-time single-channel dereverberation and separation with time-domain audio separation network. In: ISCA Interspeech (2018)
18. Schwartz, B., Gannot, S., Habets, E.A.P.: Online speech dereverberation using Kalman filter and EM algorithm. IEEE/ACM Trans. Audio, Speech, Language Proc. **23**(2) (2015)
19. Braun, S., Habets, E.A.P.: Online dereverberation for dynamic scenarios using a Kalman filter with an autoregressive model. IEEE Signal Proc. Letters **23**(12) (2016)
20. Yoshioka, T., Tachibana, H., Nakatani, T., Miyoshi, M.: Adaptive dereverberation of speech signals with speaker-position change detection. In: IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP) (2009)
21. Heymann, J., Drude, L., Haeb-Umbach, R., Kinoshita, K., Nakatani, T.: Frame-online DNN-WPE dereverberation. International Workshop on Acoustic Signal Enhancement (2018)
22. Chang, X., Zhang, W., Qian, Y., Roux, J.L., Watanabe, S.: MIMO-speech: End-to-end multi-channel multi-speaker speech recognition. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (2019)
23. Ochiai, T., Watanabe, S., Hori, T., Hershey, J.R., Xiao, X.: Unified architecture for multichannel end-to-end speech recognition with neural beamforming. IEEE Journal of Selected Topics in Signal Proc. **11**(8) (2017)
24. Zhang, W., Boeddeker, C., Watanabe, S., Nakatani, T., Delcroix, M., Kinoshita, K., Ochiai, T., Kamo, N., Haeb-Umbach, R., Qian, Y.: End-to-end dereverberation, beamforming, and speech recognition with improved numerical stability and advanced frontend. In: IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP) (2021)
25. Heymann, J., Drude, L., Haeb-Umbach, R., Kinoshita, K., Nakatani, T.: Joint optimization of neural network-based WPE dereverberation and acoustic model for robust online ASR. In: IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP) (2019)
26. Lemercier, J.-M., Thiemann, J., Konig, R., Gerkmann, T.: Customizable end-to-end optimization of online neural network-supported dereverberation for hearing devices. In: IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP) (2022)
27. Hu, Y., Kokkinakis, K.: Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners. The Journal of the Acoustical Society of America **135**, 22–8 (2014)
28. Wang, Z.-Q., Wang, D.: Deep learning based target cancellation for speech dereverberation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28** (2020)
29. Drude, L., Böddeker, C., Heymann, J., Haeb-Umbach, R., Kinoshita, K., Delcroix, M., Nakatani, T.: Integrating neural network based beamforming and weighted prediction error dereverberation. In: ISCA Interspeech (2018)
30. Maciejewski, M., Wichern, G., McQuinn, E., Roux, J.L.: WHAMR!: Noisy and reverberant single-channel speech separation. In: IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP) (2020)
31. Cohen, I.: Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. IEEE Signal Processing Letters **9**(4) (2002)

32. Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B.-H.: Speech dereverberation based on variance-normalized delayed linear prediction. IEEE Trans. Audio, Speech, Language Proc. **18**(7) (2010)
33. Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Audio, Speech, Language Proc. **33**(2) (1985)
34. Breithaupt, C., Krawczyk, M., Martin, R.: Parameterized mmse spectral magnitude estimation for the enhancement of noisy speech. In: IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP) (2008)
35. Ernst, O., Chazan, S.E., Gannot, S., Goldberger, J.: Speech dereverberation using fully convolutional networks. In: Proc. Euro. Signal Proc. Conf. (EUSIPCO) (2019)
36. Paul, D.B., Baker, J.M.: The design for the Wall Street Journal-based CSR corpus. In: Proceedings of the Workshop on Speech and Natural Language (1992). doi:10.3115/1075527.1075614
37. Wendt, T., Van De Par, S., Ewert, S.D.: A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation. Journal of the Audio Engineering Society **62**(11) (2014)
38. Thiemann, J., van de Pars, S.: A multiple model high-resolution head-related impulse response database for aided and unaided ears. EURASIP Journal on Advances in Signal Proc. **2019** (2019)
39. Kuttruff, H.: Room acoustics. CRC Press (2016)
40. Carbajal, G., Serizel, R., Vincent, E., Humbert, E.: Joint NN-supported multichannel reduction of acoustic echo, reverberation and noise. IEEE/ACM Trans. Audio, Speech, Language Proc. **28** (2020)
41. Vincent, E., Gribonval, R., Fevotte, C.: Performance measurement in blind audio source separation. IEEE/ACM Trans. Audio, Speech, Language Proc. **14**(4) (2006)
42. Li, A., Liu, W., Luo, X., Yu, G., Zheng, C., Li, X.: A simultaneous denoising and dereverberation framework with target decoupling. In: ISCA Interspeech (2021)
43. Li, A., Liu, W., Luo, X., Zheng, C., Li, X.: Decoupling magnitude and phase optimization with a two-stage deep network. In: IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP) (2021)
44. Reddy, C.K.A., Dubey, H., Gopal, V., Cutler, R., Braun, S., Gamper, H., Aichner, R., Srinivasan, S.: ICASSP 2021 Deep Noise Suppression challenge. In: IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP) (2021)

3.2 Deep Subband Filtering for Speech Dereverberation [P4]

Abstract

In this paper, we present a scheme for extending deep neural network-based multiplicative maskers to deep subband filters for speech restoration in the time-frequency domain. The resulting method can be generically applied to any deep neural network providing masks in the time-frequency domain, while requiring only few more trainable parameters and a computational overhead that is negligible for state-of-the-art neural networks. We demonstrate that the resulting deep subband filtering scheme outperforms multiplicative masking for dereverberation, while leaving the denoising performance virtually the same. We argue that this is because deep subband filtering in the time-frequency domain fits the subband approximation often assumed in the dereverberation literature, whereas multiplicative masking corresponds to the narrowband approximation generally employed for denoising.

Reference

Jean-Marie Lemercier, Julian Tobergte and Timo Gerkmann "Extending DNN-based Multiplicative Masking to Deep Subband Filtering for Improved Dereverberation", *ISCA Interspeech*, 2023. DOI: 10.21437/Interspeech.2023-1429

Copyright Notice

Authors' Contributions

Jean-Marie Lemercier is the first author of this publication. He proposed the original idea of extending time-frequency masking to deep filters for dereverberation. He conducted the dereverberation experiments and the evaluation, he wrote the manuscript. Julian Tobergte implemented the deep filtering extension and conducted the denoising experiments. He wrote his M.Sc. thesis under the supervision of Jean-Marie Lemercier, which this paper is an output of. Timo Gerkmann brought insights on the experimental validation, mathematical derivations, and reviewed the manuscript.

Extending DNN-based Multiplicative Masking to Deep Subband Filtering for Improved Dereverberation

Jean-Marie Lemercier, Julian Tobergte, Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany

{firstname.lastname}@uni-hamburg.de

Abstract

In this paper, we present a scheme for extending deep neural network-based multiplicative maskers to deep subband filters for speech restoration in the time-frequency domain. The resulting method can be generically applied to any deep neural network providing masks in the time-frequency domain, while requiring only few more trainable parameters and a computational overhead that is negligible for state-of-the-art neural networks. We demonstrate that the resulting deep subband filtering scheme outperforms multiplicative masking for dereverberation, while leaving the denoising performance virtually the same. We argue that this is because deep subband filtering in the time-frequency domain fits the subband approximation often assumed in the dereverberation literature, whereas multiplicative masking corresponds to the narrowband approximation generally employed for denoising.

Index Terms: multi-frame filtering, subband approximation, dereverberation, denoising, neural network

1. Introduction

In modern communication devices, recorded speech is corrupted when clean speech sources are affected by interfering speakers, background noise and room acoustics. Speech restoration aims to recover clean speech from the corrupted signal, whereby two distinct tasks, denoising and dereverberation, are considered here [1, 2].

Traditional speech restoration algorithms are based on statistical methods, exploiting properties of the target and interfering signals to discriminate between them [3]. These include linear prediction [4], spectral enhancement [5], inverse filtering [6], and cepstral processing [7]. Modern approaches rely mostly on machine learning. In this field, predictive methods, learning a one-to-one mapping between corrupted and clean speech through a deep neural network (DNN), are most popular [8, 9]. A large portion of DNNs used in speech restoration are trained for mask estimation, i.e. they learn a mask value to be applied to each single bin of the signal, either in a learnt domain [10] or in the time-frequency (TF) domain [11, 12]. On the opposite, some approaches employ deep filtering [13], which means that their final stage involves a convolution between the input signal and a learnt multi-frame TF filter [14, 15, 16, 17, 18, 19, 20]. In [17], this filter is parameterized as a multi-frame MVDR [21] for denoising. A DNN-parameterized weighted prediction error subband filter is proposed in [19, 18, 20]. A deep filter can also be directly learnt, e.g. in [15] as a frequency-independent time filter or in [13, 14] as a joint time-frequency filter.

This work has been funded by the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380. The authors are responsible for the content of this paper.

In this paper, we propose a deep subband filtering extension (DSFE) scheme to transform masking-based speech restoration DNNs into deep subband filters. The proposed extension is implemented by using a learnable temporal convolution at the output of the original masking DNN backbone and training the resulting architecture in an end-to-end fashion in the TF domain. Most of the time, the original masking DNN already handles multi-frame filtering internally through e.g. temporal convolutions. However, we show that enforcing explicit multi-frame subband filtering as the final stage of processing results in a significant performance increase for dereverberation while leaving the denoising performance virtually unaltered. We justify our approach by relating time-frequency multiplicative masking and deep subband filtering to the noising and reverberation corruption models respectively. The proposed approach has a negligible computational overhead and constitutes a generic module that can be plugged in any masking-based system.

The remainder of this paper is organized as follows. We first present an overview of the signal model and prerequisite assumptions for reverberation and noising corruptions. Then, we introduce our deep subband filtering extension scheme. We proceed with describing our experimental setup including data generation and training configuration. Finally we present and discuss our results.

2. Signal model

2.1. Narrowband and subband filtering

Filtering in the time-domain is obtained via convolution of a filter w with the speech signal s, yielding the filtered signal x:

$$x_t = \sum_\tau w_\tau s_{t-\tau}, \tag{1}$$

where t is the time index. A well-known result of Fourier theory is that, when transposed in the Fourier domain, such a filtering process can be expressed as a multiplication of the Fourier spectra. When using the short-time Fourier transform (STFT) however, the window used for analysis is of limited size, and spectral leakage between frequency bands can occur. Consequently, the true filtering model is:

$$\mathbf{x}_{t,f} = \sum_\tau \sum_\nu \tilde{\mathbf{w}}_{\tau,f,\nu} \mathbf{s}_{t-\tau,\nu}, \tag{2}$$

where f is the frequency index, $\mathbf{x} := \mathrm{STFT}(x)$, $\mathbf{s} := \mathrm{STFT}(s)$ and $\tilde{\mathbf{w}}_{\tau,f,\nu}$ is interpreted as a response to a time-frequency impulse $\delta_{\tau,f-\nu}$ [22]. The sum over index ν represents cross-band filtering, and the sum over index τ is a convolution along the time dimension.

The *subband approximation* ignores the effects of spectral

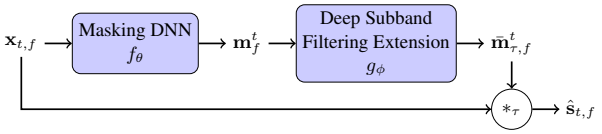


Figure 1: *Proposed model diagram. The blue blocks are learnable neural networks.*

leakage. Therefore, cross-band filtering is discarded and a single convolution is computed along the time-dimension in each frequency band independently:

$$\mathbf{x}_{t,f} = \sum_\tau \mathbf{w}_{\tau,f}\mathbf{s}_{t-\tau,f}, \tag{3}$$

where $\mathbf{w} := \mathrm{STFT}(w)$.

The *narrowband approximation* further assumes that the length of the filter w is inferior to the STFT window length, therefore zeroing out the filter taps $\{\mathbf{w}_{\tau,f}; \tau \geq 1\}$ and yielding the following filtering model :

$$\mathbf{x}_{t,f} = \mathbf{w}_f \mathbf{s}_{t,f}. \tag{4}$$

2.2. Corruption models

Speech denoising consists in removing additive background noise n from the mixture x. The forward corruption process can naturally be represented in the STFT domain by addition of the clean speech and noise spectrograms:

$$\mathbf{x} = \mathbf{n} + \mathbf{s}. \tag{5}$$

Many speech denoising approaches use time-frequency masking, i.e. they compute a mask $\mathbf{m}$ for each time-frequency bin and apply it to retrieve the clean speech estimate $\hat{\mathbf{s}}$:

$$\hat{\mathbf{s}}_{t,f} = (\mathbf{m} \odot \mathbf{x})_{t,f} := \mathbf{m}_f^t \mathbf{x}_{t,f}. \tag{6}$$

This model is similar to the narrowband approximation (4) with a *time-dependent* filter $\mathbf{m}$. We put the index t as superscript to avoid confusion with the time-convolution index τ.

In contrast to denoising, speech dereverberation aims to recover the anechoic speech corrupted by room acoustics. The signal model is exactly the filtering process in (1), where the filter w is called the room impulse response (RIR). Since the RIR length is almost always larger than the STFT window length, one cannot use the narrowband approximation (4) and has to resort to the subband approximation (3) instead. Consequently, some speech dereverberation methods perform inverse filtering in the STFT domain using the subband approximation [14, 4, 19, 18, 20]. That is, they try and estimate a filter $\bar{\mathbf{m}}$ supposed to represent the inverse of the RIR, such that the anechoic speech estimate is retrieved as:

$$\hat{\mathbf{s}}_{t,f} = (\bar{\mathbf{m}} * \mathbf{x})_{t,f} := \sum_\tau \bar{\mathbf{m}}_{\tau,f}^t \mathbf{x}_{t-\tau,f}, \tag{7}$$

with $*$ representing a convolution over the time-axis. Please note that in the model above in contrast to (3), the filter $\bar{\mathbf{m}}$ is considered *time-dependent*, same as in the time-frequency masking case. This is often assumed in order to account for non-stationarity of the RIR and estimation errors [23, 18, 20, 24].

3. Deep subband filtering extension

Many neural network-based schemes use time-frequency masking, without examining the nature of the corruption. In this section, we present our DSFE scheme, which turns the time-frequency masks produced by such DNNs into subband filters. Let f_θ be a DNN providing a mask $\mathbf{m} = f_\theta(\mathbf{x})$ in the complex spectrogram domain, such that the clean spectrogram estimate is obtained via time-frequency masking (6).

We wish to extend the mask $\mathbf{m}$ into a filter $\bar{\mathbf{m}}$ implemented by the neural network combination $\bar{\mathbf{m}} = g_\phi(f_\theta(\mathbf{x}))$, such that the clean estimate is obtained via subband filtering (7). Essentially, we want to turn masking DNNs into deep subband filters [13]. To this end, we design the *deep subband filtering extension* g_ϕ as a point-wise two-dimensional convolutional layer with tanh activation. The maps are of size $T \times F$, the kernels of size 1×1 and there are 2 input and $2N_f$ output channels corresponding to the single-frame mask and multi-frame filter real and imaginary parts, respectively:

$$g_\phi : \mathbf{m} \to \bar{\mathbf{m}} = \frac{1}{N_f} \tanh\left(\mathrm{Conv2D}(\mathbf{m}; \phi)\right). \tag{8}$$

Note that we feed the spectrogram $\mathbf{x}$ to the neural network and only use the multi-frame representation $\{\mathbf{x}_{t-\tau,f}; \tau \in [0, 1, ..., N_f - 1]\}$ for filtering. This is because the multi-frame representation does not add any relevant information with respect to $\mathbf{x}$: since most DNNs compute correlations along the time-dimension already, it is redundant to provide a vector which explicitly encodes that time-delayed information. The proposed algorithm is summarized on Figure 1.

As the corresponding inverse filtering model better fits the corruption model for reverberation, we expect our DSFE method to perform better at dereverberation than its masking counterpart, and not produce significant changes for denoising.

4. Experimental setup

4.1. Data

Both datasets for denoising and dereverberation experiments use the WSJ0 corpus [25] for clean speech sources. The training, validation and test splits comprise 101, 10 and 8 speakers for a total of 12777, 1206 and 651 utterances and a length of 25, 2.3 and 1.5 hours of speech respectively, sampled at 16 kHz.

Speech Denoising: The WSJ0+Chime dataset is generated using clean speech extracts from the WSJ0 corpus and noise signals from the CHiME3 dataset [26]. The mixture signal is created by randomly selecting a noise file and adding it to a clean utterance with a signal-to-noise ratio (SNR) sampled uniformly between -6 and 14 dB.

Speech Dereverberation: The WSJ0+Reverb dataset is generated using clean speech data from the WSJ0 corpus and convolving each utterance with a simulated RIR. We use the `pyroomacoustics` library [27] to simulate the RIRs. The reverberant room is modeled by sampling uniformly a target T_{60} between 0.4 and 1.0 seconds and room length, width and height in [5,15]×[5,15]×[2,6] m. The anechoic target is generated using the T_{60}-shortening method [28], where the RIR is shaped by a decaying exponential window so that the resulting T_{60} equals 200ms. This results in an average direct-to-reverberation ratio (DRR) of -5.3 dB.

4.2. Single-frame DNN backbone

In this paper, we use the GaGNet architecture by [11], a state-of-the-art denoising neural network, which is the successor of [29] which ranked first in the real-time enhancement track of the DNS-2021 challenge. GaGNet leverages magnitude-only and complex-domain information in parallel with temporal convolutional networks. The rationale is to obtain a coarse estimation with the magnitude-processing *glance* modules, and to refine this estimation with *gaze* modules processing the real and imaginary parts of the complex spectrogram. Between each repeated glance and gaze module, an approximate complex ratio mask [12] is applied on the current version of the signal to enforce a coherent filtering process and stabilize training. Finally, the network outputs multiplicative mask values for the real and imaginary parts. We name our proposed method *DSFE-GaGNet*, which is the concatenation of GaGNet with the DSFE module g_ϕ. Although we focus on GaGNet in this work, please note that our DSFE method is compatible with any architecture performing mask estimation in the complex STFT domain. It could even be envisaged to use a similar extension in a different domain e.g. learnt by an DNN encoder.

4.3. Training configuration

We use the same training configuration as GaGNet [11]: the STFT uses a Hann window with 320 points and 50% overlap at a sample rate of 16 kHz. We employ square-root compression on the magnitude spectrogram. Therefore, the features that are fed to GaGNet are: $\mathrm{cat}(\sqrt{|\mathbf{x}|}\cos(\phi_x), \sqrt{|\mathbf{x}|}\sin(\phi_x))$, where $\mathbf{x} = |\mathbf{x}|\exp(j\phi_x)$ is the noisy complex spectrogram. The training loss is a sum of mean square errors with respect to the real part, imaginary part and magnitude of the clean and estimated spectrograms. The networks are trained with the Adam optimizer with a learning rate of 0.0005. Contrarily to [11], we use mini-batches of size 48 and use early stopping with a patience of 50 epochs and a maximum of 2000 epochs.

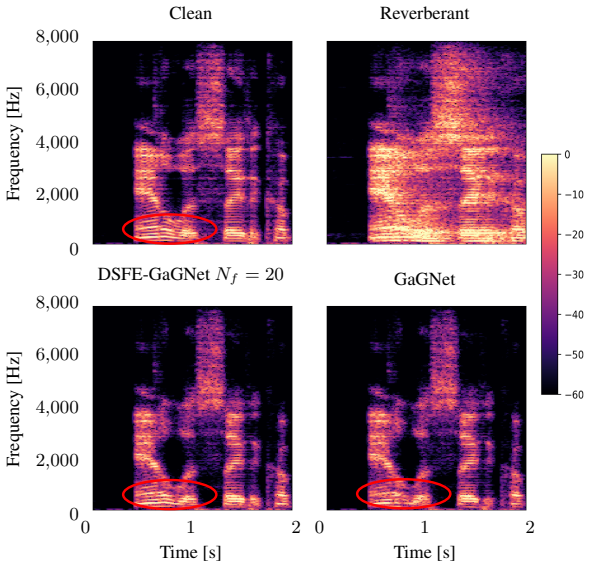


Figure 2: *Log-energy spectrograms of clean, reverberant and processed signals form the WSJ0+Reverb dataset. The harmonic structure in the red circle is altered with GaGNet and better preserved with DSFE-GaGNet.* $\mathrm{T}_{60} = 0.85$s.

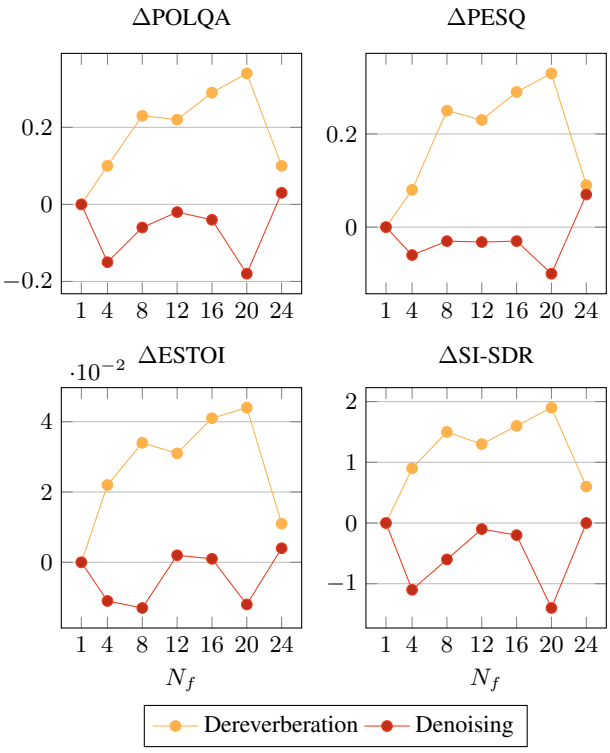


Figure 3: *Instrumental metrics improvements of DSFE-GaGNet with respect to single-frame GaGNet for speech denoising on WSJ0+Chime and dereverberation as a function of the number of frames* N_f.

4.4. Evaluation

We conduct instrumental evaluation using classical speech metrics like Perceptual Objectve Listening Quality Analysis (POLQA) [30], Perceptual Evaluation of Speech Quality (PESQ) [31], Extended Short-Term Objective Intelligibility (ESTOI) [32] as well as scale-invariant (SI-) signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and signal-to-artifacts ratio (SAR) [33]. We also report the number of million single-point floating operations per second of processed speech (MFLOPS$\cdot$s^{-1}) as provided by the `pypapi` library[1].

5. Experimental results and discussion

5.1. Multi-frame filtering for speech enhancement tasks

We report results for dereverberation on WSJ0+Reverb and denoising on WSJ0+Chime in tables 1 and 2 respectively. For a more direct comparison, we group these experiments in Figure 3 by showing the improvements of our method DSFE-GaGNet, with respect to its single-frame GaGNet counterpart as a function of N_f for both dereverberation and denoising.

For dereverberation, we observe a monotonic increase in all instrumental metrics as more frames are used in DSFE-GaGNet. The performance peaks at $N_f = 20$ with an improvement of .33 PESQ, .04 ESTOI and 1.9dB SI-SDR over the single-frame baseline. This improvement then decreases, as we observe that training is less stable with a high number of frames e.g. $N_f = 24$. We observe on the spectrograms displayed in Figure 2 that

[1] https://github.com/flozz/pypapi

Table 1: *Dereverberation results obtained on the WSJ0-Reverb dataset. Values indicate mean and standard deviation.*

Method	N_f	POLQA	PESQ	ESTOI	SI-SDR	SI-SIR	SI-SAR	MFLOPS·s^{-1}
Mixture	–	1.94 ± 0.40	1.51 ± 0.30	0.62 ± 0.12	1.2 ± 2.8	-0.8 ± 2.5	–	–
GaGNet	1	3.07 ± 0.43	2.52 ± 0.44	0.83 ± 0.06	6.0 ± 2.4	5.9 ± 2.6	6.0 ± 2.4	367.2
DSFE-GaGNet	4	3.17 ± 0.41	2.60 ± 0.44	0.85 ± 0.05	6.9 ± 2.3	6.8 ± 2.7	6.4 ± 2.2	368.6
DSFE-GaGNet	8	3.30 ± 0.42	2.77 ± 0.44	0.86 ± 0.05	7.5 ± 2.1	7.4 ± 2.7	6.7 ± 2.1	368.8
DSFE-GaGNet	12	3.29 ± 0.42	2.75 ± 0.44	0.86 ± 0.05	7.3 ± 2.2	7.1 ± 2.7	6.7 ± 2.2	370.0
DSFE-GaGNet	16	3.36 ± 0.42	2.81 ± 0.45	0.87 ± 0.05	7.6 ± 2.2	7.6 ± 2.7	6.7 ± 2.2	370.3
DSFE-GaGNet	20	$\mathbf{3.41 \pm 0.40}$	$\mathbf{2.85 \pm 0.44}$	$\mathbf{0.87 \pm 0.05}$	$\mathbf{7.9 \pm 2.3}$	$\mathbf{7.9 \pm 2.9}$	$\mathbf{6.9 \pm 2.2}$	371.6
DSFE-GaGNet	24	3.17 ± 0.43	2.61 ± 0.45	0.84 ± 0.06	6.6 ± 2.4	6.5 ± 2.7	6.2 ± 2.3	371.8

Table 2: *Denoising results obtained on the WSJ0+Chime dataset. Values indicate mean and standard deviation.*

Method	N_f	POLQA	PESQ	ESTOI	SI-SDR	SI-SIR	SI-SAR	MFLOPS·s^{-1}
Mixture	–	2.08 ± 0.64	1.38 ± 0.32	0.65 ± 0.18	4.3 ± 5.8	4.3 ± 5.8	–	–
GaGNet	1	3.48 ± 0.60	2.75 ± 0.59	$\mathbf{0.89 \pm 0.08}$	$\mathbf{15.5 \pm 4.1}$	26.1 ± 4.5	$\mathbf{16.0 \pm 4.3}$	367.2
DSFE-GaGNet	4	3.33 ± 0.64	2.69 ± 0.59	0.88 ± 0.08	14.4 ± 4.1	25.2 ± 5.0	14.8 ± 4.2	368.6
DSFE-GaGNet	8	3.42 ± 0.63	2.72 ± 0.60	0.88 ± 0.08	14.9 ± 4.1	26.2 ± 4.8	15.4 ± 4.2	368.8
DSFE-GaGNet	12	3.46 ± 0.61	2.75 ± 0.58	0.89 ± 0.08	15.0 ± 4.0	27.0 ± 5.1	15.4 ± 4.0	370.0
DSFE-GaGNet	16	3.44 ± 0.63	2.72 ± 0.59	0.89 ± 0.08	15.3 ± 4.2	26.7 ± 4.8	15.7 ± 4.3	370.3
DSFE-GaGNet	20	3.30 ± 0.63	2.65 ± 0.57	0.88 ± 0.08	14.1 ± 3.9	24.7 ± 5.0	14.6 ± 3.9	371.6
DSFE-GaGNet	24	$\mathbf{3.51 \pm 0.60}$	$\mathbf{2.82 \pm 0.56}$	$\mathbf{0.89 \pm 0.08}$	$\mathbf{15.5 \pm 4.1}$	$\mathbf{27.1 \pm 4.8}$	$\mathbf{16.0 \pm 4.2}$	371.8

DSFE-GaGNet preserves the harmonic structure in some cases where that structure is altered by GaGNet.

In the denoising case, the DSFE module reveals useless as DSFE-GaGNet performance saturates at the level of the single-frame GaGNet, or even worsens with more frames, at the exception of $N_f = 24$ where marginal improvements are observed.

This comparison suggests that subband filtering should be adopted when it fits the corruption model, i.e. for convolutive signal models like reverberation where the narrowband approximation requirements are not satisfied. In that case, we can obtain remarkable improvements at a very low computational cost: our best model DSFE-GaGNet with $N_f = 20$ only requires 4.4 MFLOPS·s^{-1} more than GaGNet, that is, a relative 1.2% increase. Furthermore, the temporal convolution used in the DSFE module with $N_f = 20$ frames employs only 96 trainable parameters, which is negligible compared to the 5.9M parameters of the original GaGNet backbone. Finally, since DSFE-GaGNet only uses past frames, the algorithmic latency does not increase and is still dominated by the length of the STFT synthesis window i.e. 20ms.

5.2. Ablation study

In Table 3 we present results of an ablation study showing various training strategies for DSFE-GaGNet. The default training configuration is denoted as *Join*, i.e. when both the DSFE module and the GaGNet backbone are trained jointly from scratch. We also try pretraining the GaGNet backbone and subsequently tuning the DSFE module parameters, either leaving the GaGNet backbone frozen (*Pretrain+Freeze*) or finetuning it along the DSFE parameters (*Pretrain+Finetune*). As expected, joint training performs best, but the improvement over *Pretrain+Finetune* is marginal. This highlights that it is paramount to jointly tune the DSFE parameters along with the single-frame backbone, at least at some stage of the training.

Table 3: *Dereverberation results of DSFE-GaGNet on WSJ0+Reverb. All approaches use $N_f = 20$ frames. Values indicate mean and standard deviation.*

Strategy	POLQA	ESTOI	SI-SDR
Mixture	1.94 ± 0.40	0.62 ± 0.12	1.2 ± 2.8
Pretrain+Freeze	3.19 ± 0.42	0.84 ± 0.06	6.8 ± 2.4
Pretrain+Finetune	3.40 ± 0.41	0.86 ± 0.05	7.5 ± 2.5
Join	$\mathbf{3.41 \pm 0.40}$	$\mathbf{0.87 \pm 0.05}$	$\mathbf{7.9 \pm 2.3}$

6. Conclusion

We present a deep subband filtering extension scheme transforming DNNs performing time-frequency multiplicative masking into deep subband filters. We show that such an extension fits the subband filtering approximation used for dereverberation in the STFT domain, while time-frequency masking fits the narrowband filtering approximation used for denoising. Consequently, we show that our deep subband filtering extension significantly increases dereverberation performance while leaving denoising performance virtually the same. The proposed extension scheme can be generically applied to any DNN baseline performing time-frequency masking, with an insignificant increase in inference time and model capacity. Ablation studies suggest that the deep subband filtering extension module should be trained jointly with the original single-frame DNN, at least at some stage of the training.

7. References

[1] P. Naylor and N. Gaubitch, *Speech Dereverberation*, vol. 59. Springer, Jan. 2011.

[2] S. J. Godsill, P. J. W. Rayner, and O. Cappé, *Digital audio restoration*. Springer, Sept. 1998.

[3] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement* (E. Vincent, ed.), John Wiley & Sons, 2018.

[4] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, (Las Vegas, USA), May 2008.

[5] E. Habets, *Single- and Multi-Microphone Speech Dereverberation Using Spectral Enhancement*. PhD thesis, Jan. 2007.

[6] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, (Florence, Italy), May 2014.

[7] T. Gerkmann, "Cepstral weighting for speech dereverberation without musical noise," in *2011 19th European Signal Processing Conference*, (Barcelona, Spain), Sept. 2011.

[8] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[9] 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.

[10] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[11] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, vol. 187, p. 108499, 2022.

[12] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 25, no. 7, pp. 1492–1501, 2017.

[13] W. Mack and E. A. P. Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Proc. Letters*, vol. 27, pp. 61–65, 2020.

[14] H. Schröter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, "DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, (Singapore, Singapore), May 2022.

[15] H. Schröter, T. Rosenkranz, A. N. Escalante-B, M. Aubreville, and A. Maier, "Clcnet: Deep learning-based noise reduction for hearing aids using complex linear coding," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, pp. 6949–6953, 2020.

[16] S. Lv, Y. Hu, S. Zhang, and L. Xie, "DCCRN+: Channel-wise subband DCCRN with SNR estimation for speech enhancement," in *Interspeech*, (Brno, Czech Republic), Sept. 2021.

[17] 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), May 2021.

[18] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Frame-online DNN-WPE dereverberation," in *Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, (Tokyo, Japan), Sept. 2018.

[19] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation," in *Interspeech*, (Stockholm, Sweden), Sept. 2017.

[20] J.-M. Lemercier, J. Thiemann, R. Koning, and T. Gerkmann, "Customizable end-to-end optimization of online neural network-supported dereverberation for hearing devices," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, (Singapore, Singapore), May 2022.

[21] 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.

[22] Y. Avargel and I. Cohen, "System Identification in the Short-Time Fourier Transform Domain With Crossband Filtering," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 15, pp. 1305–1319, May 2007.

[23] A. Jukić, T. van Waterschoot, and S. Doclo, "Adaptive speech dereverberation using constrained sparse multichannel linear prediction," *IEEE Signal Processing Letters*, vol. 24, no. 1, pp. 101–105, 2017.

[24] J.-M. Lemercier, J. Thiemann, R. Koning, and T. Gerkmann, "Neural Network-augmented Kalman Filtering for Robust Online Speech Dereverberation in Noisy Reverberant Environments," in *Interspeech*, (Incheon, South Korea), Sept. 2022.

[25] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete," *Linguistic Data Consortium*, May 2007.

[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)*, (Scottsdale, USA), Dec. 2015.

[27] I. D. R. Scheibler, E. Bezzam, "Pyroomacoustics: A Python package for audio room simulations and array processing algorithms," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, (Calgary, Canada), Apr. 2018.

[28] R. Zhou, W. Zhu, and X. Li, "Speech dereverberation with a reverberation time shortening target," *arXiv*, 2022.

[29] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "ICASSP 2021 Deep Noise Suppression Challenge: Decoupling magnitude and phase optimization with a two-stage deep network," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, (Brno, Czech Republic), June 2021.

[30] ITU-T Rec. P.863, "Perceptual objective listening quality prediction," *Int. Telecom. Union (ITU)*, 2018.

[31] 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)*, (Salt Lake City, USA), May 2001.

[32] 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.

[33] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, (Brighton, United Kingdom), May 2019.

4

Supervised Conditional Diffusion Models for Speech Dereverberation

4.1 Diffusion Models for Audio Restoration [P5]

Abstract

With the development of audio playback devices and fast data transmission, the demand for high sound quality is rising for both entertainment and communications. In this quest for better sound quality, challenges emerge from distortions and interferences originating at the recording side or caused by an imperfect transmission pipeline. To address this problem, audio restoration methods aim to recover clean sound signals from the corrupted input data. We present here audio restoration algorithms based on diffusion models, with a focus on speech enhancement and music restoration tasks. Traditional approaches, often grounded in handcrafted rules and statistical heuristics, have shaped our understanding of audio signals. In the past decades, there has been a notable shift towards data-driven methods that exploit the modeling capabilities of DNNs. Deep generative models, and among them diffusion models, have emerged as powerful techniques for learning complex data distributions. However, relying solely on DNN-based learning approaches carries the risk of reducing interpretability, particularly when employing end-to-end models. Nonetheless, data-driven approaches allow more flexibility in comparison to statistical model-based frameworks, whose performance depends on distributional and statistical assumptions that can be difficult to guarantee. Here, we aim to show that diffusion models can combine the best of both worlds and offer the opportunity to design audio restoration algorithms with a good degree of interpretability and a remarkable performance in terms of sound quality. We explain the diffusion formalism and its application to the conditional generation of clean audio signals. We believe that diffusion models open an exciting field of research with the potential to spawn new audio restoration algorithms that are natural-sounding and remain robust in difficult acoustic situations.

Reference

Jean-Marie Lemercier, Julius Richter, Simon Welker, Eloi Moliner, Vesa Välimäki and Timo Gerkmann, "Diffusion Models for Audio Restoration", *IEEE Signal Processing Magazine*, arXiv: https://arxiv.org/abs/2402.09821

Copyright Notice

Authors' Contributions

Jean-Marie Lemercier is the first author, he wrote the last section "Practical requirements of Diffusion-based Sampling for Audio Tasks" and the paragraph "Model-based processing with diffusion models" of the section "Basics of Diffusion Models". Julius Richter wrote the abstract and first versions of the introduction and the section "Basics of Diffusion Models". Julius Richter and Jean-Marie Lemercier co-lead the overall writing process, and revised the manuscript. Simon Welker wrote the section "Conditional Generation with Diffusion Models". Eloi Moliner wrote the section "Diffusion Models for Inverse Problems". Timo Gerkmann and Vesa Välimäki brought insights into the organization of the writing and the structure of the manuscript. All authors contributed to reviewing and revising the manuscript.

Diffusion Models for Audio Restoration

Invited paper for the SPM Special entitled “Model-based and Data-Driven Audio Signal Processing”.

Jean-Marie Lemercier[†], Julius Richter[†], Simon Welker[†], Eloi Moliner[°], Vesa Välimäki[°], Timo Gerkmann[†]

[†]Signal Processing (SP), Department of Informatics, Universität Hamburg, Germany

[°]Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland

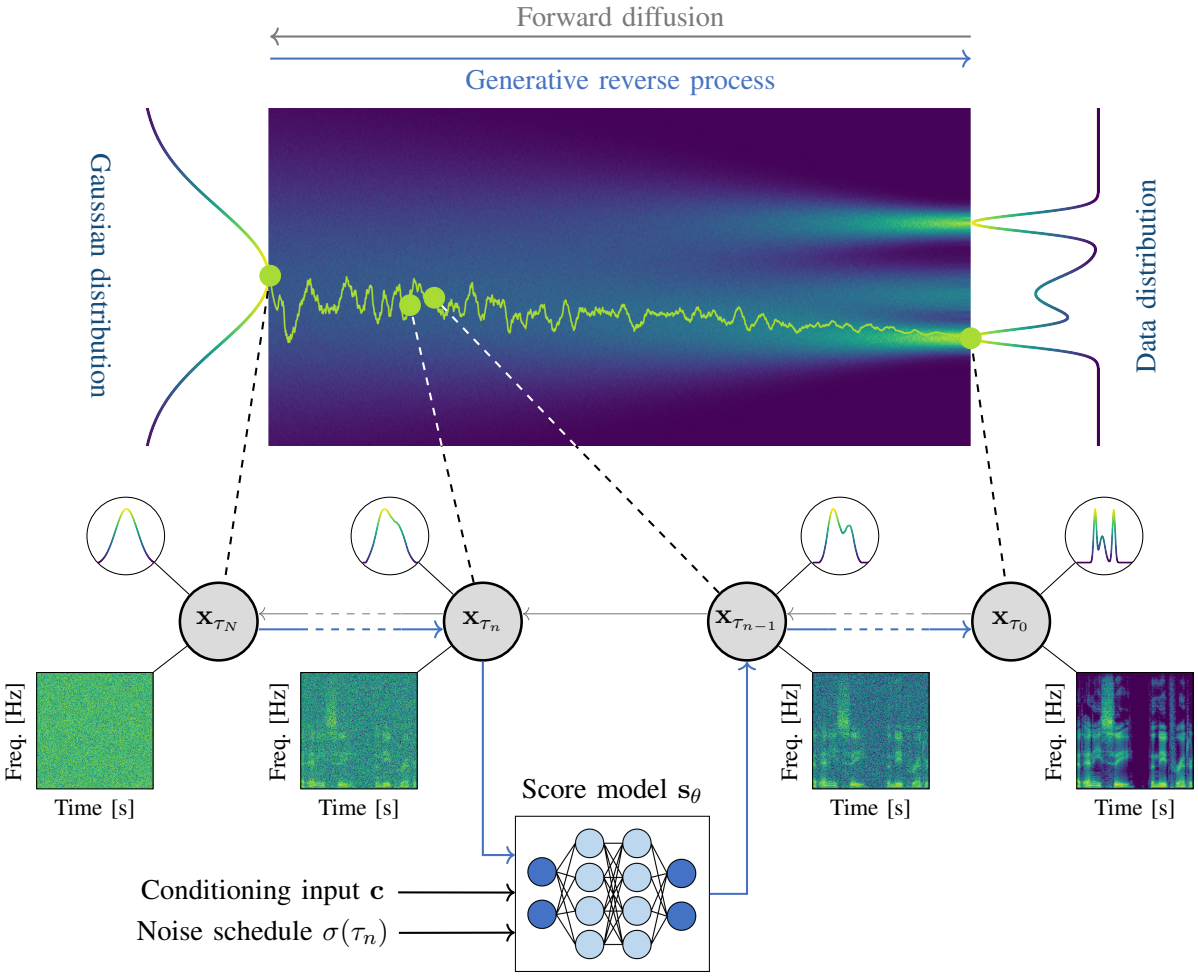


Fig. 1: A continuous-time diffusion model transforms (left) a Gaussian distribution to (right) an intractable data distribution through a stochastic process $\{\mathbf{x}_\tau\}_{\tau\in[0,T]}$ with marginal distributions $\{p_\tau(\mathbf{x}_\tau)\}_{\tau\in[0,T]}$. During training, the forward diffusion is simulated by adding Gaussian noise and rescaling the data, and a score model $\mathbf{s}_\theta$ with parameters θ learns the score function $\nabla_{\mathbf{x}_\tau}\log p_\tau(\mathbf{x}_\tau)$. During the generative reverse process, the process time τ is discretized to steps $\{\tau_0,\dots,\tau_N\}$ and followed in reverse from $\tau_N = T$ to $\tau_0 = 0$. (Bottom) The next state $\mathbf{x}_{\tau_{n-1}}$ is obtained based on the previous state $\mathbf{x}_{\tau_n}$ using an estimate given by the score model. The score model is conditioned by the noise scale at the current time step, $\sigma(\tau_n)$, and optional conditioning $\mathbf{c}$ to guide the generation such as e.g. a text description.

With the development of audio playback devices and fast data transmission, the demand for high sound quality is rising for both entertainment and communications. In this quest for better sound quality, challenges emerge from distortions and interferences originating at the recording side or caused by an imperfect transmission pipeline. To address this problem, audio restoration methods aim to recover clean sound signals from the corrupted input data. We present here audio restoration algorithms based on diffusion models, with a focus on speech enhancement and music restoration tasks.

Traditional approaches, often grounded in handcrafted rules and statistical heuristics, have shaped our understanding of audio signals. In the past decades, there has been a notable shift towards data-driven methods that exploit the modeling capabilities of deep neural networks (DNNs). Deep generative models, and among them diffusion models, have emerged as powerful techniques for learning complex data distributions. However, relying solely on DNN-based learning approaches carries the risk of reducing interpretability, particularly when employing end-to-end models. Nonetheless, data-driven approaches allow more flexibility in comparison to statistical model-based frameworks, whose performance depends on distributional and statistical assumptions that can be difficult to guarantee. Here, we aim to show that diffusion models can combine the best of both worlds and offer the opportunity to design audio restoration algorithms with a good degree of interpretability and a remarkable performance in terms of sound quality.

In this article, we review the use of diffusion models for audio restoration. We explain the diffusion formalism and its application to the conditional generation of clean audio signals. We believe that diffusion models open an exciting field of research with the potential to spawn new audio restoration algorithms that are natural-sounding and remain robust in difficult acoustic situations.

INTRODUCTION

Traditional audio restoration methods exploit statistical properties of audio signals, such as auto-regressive modeling for click removal [1] or probabilistic modeling for speech enhancement and separation [2], by using various representations like time-domain waveforms, spectrograms, or cepstra. Although they are robust to many scenarios, such methods struggle with highly non-stationary sources or interferences that appear in real-life scenarios. In the past decade, audio signal processing algorithms have benefited greatly from the introduction of data-driven approaches based on DNNs [3]. Among these methods, a broad class leverages *predictive models* that learn to map a given input to a desired output. Note that the term *predictive models* covers both classification and regression tasks, unlike *discriminative models* [4]. In a typical supervised setting, a predictive model is trained on a labeled dataset to minimize a certain point-wise loss function between the processed input and the clean target. Following the principle of empirical risk minimization, the goal of predictive modeling is to find a model with minimal average error over the training data, where the generalization ability of the model is usually assessed on a validation set of unseen data. By employing ever-larger models and datasets—a current trend in deep learning—strong generalization

can be achieved. However, many purely data-driven approaches are considered black boxes and remain largely unexplainable and non-interpretable. Moreover, these models typically produce deterministic outputs, disregarding the inherent uncertainty in their results.

Generative models follow a different learning paradigm, namely estimating and sampling from an unknown data distribution. This can be used to infer a measure of uncertainty for their predictions and to allow the generation of multiple valid estimates instead of a single best estimate as in predictive approaches [4]. Furthermore, incorporating prior knowledge into generative models can guide the learning process and enforce desired properties about the learned distribution. generative adversarial networks (GANs) [5] and variational auto-encoders (VAEs) [6] have been instrumental in the early developments of generative models. Subsequent to these approaches, *diffusion models* [7], [8] have emerged as a distinct class of deep generative models that boast an impressive ability to learn complex data distributions such as that of natural images [7], [8], music [9], and speech [10]. Diffusion models generate data samples through iterative transformations, transitioning from a tractable prior distribution (e.g. Gaussian) to a target data distribution, as visualized in Figure 1. This iterative generation scheme is formalized as a stochastic process and is parameterized with a DNN that is trained to address a Gaussian denoising task.

From a practical point of view, diffusion models have become popular because they can generate high-quality samples while being simpler to train than GANs. Moreover, combining data-driven machine learning techniques with mathematical concepts, such as stochastic processes, opens up possibilities for modeling conditional data distributions and integrating Bayesian inference tools. In audio processing, this has spawned new types of algorithms that adopt diffusion models for restoration tasks such as speech enhancement [11], [12] or music restoration [9]. Here, we present a comprehensive overview and categorization of novel techniques for solving audio restoration problems using diffusion models in a data-driven, model-based fashion.

In the following, we first look at the basics of diffusion models and show how they can be used for model-based processing. We then examine conditional generation with diffusion models for audio restoration tasks, distinguishing between three different conditioning techniques. In particular, we look at diffusion models for audio inverse problems with a known degradation operator and its extension to blind inverse problems when the degradation operator is unknown. We conclude by discussing the practical requirements of diffusion models for audio restoration tasks, examining sampling speed and robustness to adverse conditions.

BASICS OF DIFFUSION MODELS

With the development of DNNs and the increase in computational power, deep generative modeling has become one of the leading directions in machine learning with a variety of applications. Deep generative models aim to design a generation process for data that resembles real-world examples, e.g., natural speech produced by a human speaker. This involves modeling the probability distribution of highly structured and complex data such that learning

and sampling are computationally tractable. One way to realize generative modeling is based on the assumption that the data is generated by some random process involving unobserved variables. Such *hidden variable models* map samples from a tractable distribution, such as the Gaussian distribution, to samples that are likely to represent target data points. From this perspective of hidden variable models, we discuss diffusion models as a distinct class of deep generative models whose hidden variables are parameterized via a stochastic process.

Diffusion models break down the problem of generating high-dimensional complex data into a series of easier *denoising* tasks. Training such a denoising model first requires defining a *forward diffusion process*, which gradually adds noise to the data points of a dataset. This corruption process progressively turns the data distribution into a Gaussian distribution, as shown in Figure 1 from right to left (gray arrows). In turn, data generation is accomplished by reversing the corruption process. First, a random sample is drawn from a Gaussian distribution, and then the model iteratively removes noise from this initial point, ultimately yielding a sample from the data distribution. This *reverse diffusion process* is illustrated in Figure 1 from left to right (blue arrows).

Formally, the forward diffusion process can be represented by the Markov chain $\mathbf{x}_0 \to \mathbf{x}_1 \to \cdots \to \mathbf{x}_N$ with $\mathbf{x}_0 \in \mathbb{R}^d$ sampled from the data distribution and fixed Gaussian transition probabilities $q(\mathbf{x}_n|\mathbf{x}_{n-1})$. The resulting directed graphical model is depicted with gray circles in Figure 1. The generative model is then described by a Markov chain in reverse order $\mathbf{x}_N \to \mathbf{x}_{N-1} \to \cdots \to \mathbf{x}_0$ with $\mathbf{x}_N$ sampled from the prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. To accomplish the generation task, a DNN is trained to denoise the sample $\mathbf{x}_n$. Specifically, it learns to approximate the transition probabilities of the reverse Markov chain $p(\mathbf{x}_{n-1}|\mathbf{x}_n)$ [7].

This discrete-time Markov chain formulation of diffusion models can be generalized to continuous-time stochastic processes by letting the number of steps N grow infinitely, conversely making the distance between steps infinitely small. This facilitates the design of novel diffusion processes and allows the use of more flexible sampling schemes [8]. Specifically, the corresponding forward diffusion process is defined as a stochastic process $\{\mathbf{x}_\tau\}_{\tau\in[0,T]}$, i.e. a collection of random variables indexed by a continuous process time $\tau \in [0, T]$ [13]. The process time τ in stochastic processes intuitively corresponds to the index n in Markov chains. It is important to note that the process time τ is completely unrelated to the time dimension of the audio signal. A single random realization of the stochastic process $\{\mathbf{x}_\tau\}_{\tau\in[0,T]}$ is depicted by the green trajectory traversing Figure 1. The conditional distribution characterizing the forward diffusion model is the *transition kernel* $q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)$ instinctively related to the probability $q(\mathbf{x}_n|\mathbf{x}_0) := \prod_{i=1}^{n} q(\mathbf{x}_i|\mathbf{x}_{i-1})$ in Markov chains. The transition kernel $q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)$ can be computed by solving a stochastic differential equation (SDE), which is a differential equation where some of the coefficients are random [13]. Specifically, we define the so-called *forward SDE* as

$$\mathrm{d}\mathbf{x}_\tau = \mathbf{f}(\mathbf{x}_\tau, \tau)\,\mathrm{d}\tau + g(\tau)\,\mathrm{d}\mathbf{w}_\tau\,, \tag{1}$$

where the function $\mathbf{f} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ is referred to as the *drift coefficient* and relates to the deterministic part of the SDE. The function $g : \mathbb{R} \to \mathbb{R}$ is called the *diffusion coefficient* and controls the amount of randomness in the SDE. More precisely, the diffusion coefficient $g(\tau)$ scales the noise injected by the stochastic process $\mathbf{w}_\tau$. In most cases, $\mathbf{w}_\tau$ is chosen to be a *Wiener process*, which is a stochastic process with independent and normally distributed increments, i.e. $\mathbf{w}_{\tau+\mathrm{d}\tau} - \mathbf{w}_\tau \sim \mathcal{N}(\mathbf{0}, \mathrm{d}\tau\, \mathbf{I})$ [13]. If the drift coefficient $\mathbf{f}$ is an affine function of $\mathbf{x}_\tau$ and the diffusion coefficient g is independent of $\mathbf{x}_\tau$, then the transition kernel has a simple Gaussian form

$$q_\tau(\mathbf{x}_\tau|\mathbf{x}_0) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}_0, \tau), \sigma(\tau)^2\mathbf{I})\,, \tag{2}$$

where the mean $\boldsymbol{\mu}(\mathbf{x}_0, \tau)$ and standard deviation $\sigma(\tau)$ are obtained by analytically solving the SDE and computing the first and second moments of the solution [13].

The reverse diffusion process, i.e. the generation process, is also a stochastic process $\{\mathbf{x}_\tau\}_{\tau\in[0,T]}$ parameterized by the process time $\tau \in [0, T]$ flowing in the reverse direction, with $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \sigma(T)^2\boldsymbol{I})$. Reversing the process time axis in (1) results in another SDE called the *reverse SDE* whose marginal distributions match those of the forward SDE [8]. Therefore, denoising the sample $\mathbf{x}_\tau$ boils down to solving the reverse SDE

$$\mathrm{d}\mathbf{x}_\tau = \left[\mathbf{f}(\mathbf{x}_\tau, \tau) - g(\tau)^2\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)\right] \mathrm{d}\tau + g(\tau)\, \mathrm{d}\bar{\mathbf{w}}_\tau\,, \tag{3}$$

where $\mathrm{d}\tau < 0$ as the process axis is traveled in the reverse direction. The stochastic process $\bar{\mathbf{w}}_\tau$ is another Wiener process associated to this reverse process axis, i.e. $\bar{\mathbf{w}}_{\tau+\mathrm{d}\tau} - \bar{\mathbf{w}}_\tau \sim \mathcal{N}(\mathbf{0}, -\mathrm{d}\tau\, \mathbf{I})$. The quantity $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$ (with $\nabla_{\mathbf{x}_\tau}$ representing the gradient operator with respect to $\mathbf{x}_\tau$) is called *score function* and is a vector field informative about the variations of the process state's logarithmic probability density. The score function $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$ is generally intractable and we need to approximate it with a DNN called *score model* $\mathbf{s}_\theta(\mathbf{x}_\tau, \tau)$ with parameters θ. Vincent et al. [14] have shown that the score model $\mathbf{s}_\theta(\mathbf{x}_\tau, \tau)$ can be optimized using *denoising score matching*, i.e. matching the score of the Gaussian transition kernel $q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)$ instead of the score of the unknown probability $p_\tau(\mathbf{x}_\tau)$. The score of the transition kernel $q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)$ can be obtained from Eq. (2) as

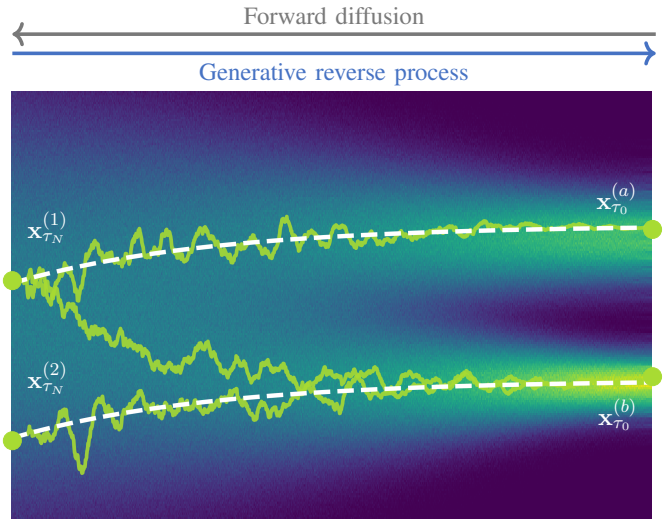
$$\nabla_{\mathbf{x}_\tau} \log q_\tau(\mathbf{x}_\tau|\mathbf{x}_0) = -\frac{\mathbf{x}_\tau - \boldsymbol{\mu}(\mathbf{x}_0, \tau)}{\sigma(\tau)^2}\,. \tag{4}$$

The score model $\mathbf{s}_\theta$ is therefore trained using the denoising score-matching objective [14]

$$\mathbb{E}_{\substack{\mathbf{x}_0\sim p(\mathbf{x}_0)\\ \tau\sim\mathcal{U}(0,T)\\ \mathbf{x}_\tau\sim q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)}} \left[\lambda(\tau) \left\| \mathbf{s}_\theta(\mathbf{x}_\tau, \tau) + \frac{\mathbf{x}_\tau - \boldsymbol{\mu}(\mathbf{x}_0, \tau)}{\sigma(\tau)^2} \right\|_2^2\right], \tag{5}$$

where a data example $\mathbf{x}_0$ is first sampled from the training set. Then, a process time τ is sampled uniformly between 0 and T, and the diffusion state $\mathbf{x}_\tau$ is obtained by sampling from the transition kernel (2). Here $\lambda(\tau)$ is a

time-dependent scaling factor, chosen empirically to stabilize training [8].



Sample initial state $\mathbf{x}_{\tau_N} \sim \mathcal{N}(\mathbf{0}, \sigma(\tau_N)^2\mathbf{I})$
for $n \in \{N, \dots, 1\}$ **do**
    Get reverse step size $\Delta\tau_n = \tau_{n-1} - \tau_n < 0$
    Obtain score estimate $\mathbf{s}_\theta(\mathbf{x}_{\tau_n}, \tau_n)$
    if *Deterministic sampler* **then**
        Take probability-flow ODE step
    else if *Stochastic sampler* **then**
        Take reverse SDE step
    Obtain next state $\mathbf{x}_{\tau_{n-1}} = \mathbf{x}_{\tau_n} + \mathrm{d}\mathbf{x}_{\tau_n}$
Output: $\mathbf{x}_0$

Fig. 2: Stochastic (green solid lines) and deterministic (white dashed lines) sampling trajectories. The stochastic sampler discretizes the reverse SDE (3) where noise is added at each sampling step by the Wiener process $\bar{\mathbf{w}}_\tau$. The deterministic sampler uses the probability-flow ordinary differential equation (ODE), which does not re-introduce noise. Two different initial points $\mathbf{x}_{\tau_N}^{(1)}$ and $\mathbf{x}_{\tau_N}^{(2)}$ are sampled, and two realizations of the stochastic sampler are shown for the same initial state $\mathbf{x}_{\tau_N}^{(1)}$. The target distribution has two modes $\mathbf{x}_{\tau_0}^{(a)}$ and $\mathbf{x}_{\tau_0}^{(b)}$.

Once the score model $\mathbf{s}_\theta$ has been trained, it allows the generation of new samples from the learned data distribution by solving the reverse SDE (3). In practice, this is done by first discretizing the process time axis into N steps $\{\tau_N, \tau_{N-1}, \dots, \tau_0\}$ with a step size $\Delta\tau_n := \tau_n - \tau_{n-1}$, often chosen uniformly. Then an initial condition $\mathbf{x}_{\tau_N}$ is sampled and the reverse SDE (3) is integrated between $\tau_N = T$ and $\tau_0 = 0$ using a numerical approximation method called SDE *solver* [15]. A differential equation solver approximates the trajectory between successive steps $\mathbf{x}_{\tau_n}$ and $\mathbf{x}_{\tau_{n-1}}$ as a piecewise polynomial function (linear if first-order solver, quadratic if second-order, etc.), whose coefficients depend on the terms in Eq. (3). An SDE solver, in particular, considers two polynomial functions (with potentially distinct degrees) to model the deterministic and stochastic terms, respectively. For instance, the widespread Euler-Maruyama method is an SDE solver with first-order polynomial approximation for both the deterministic and stochastic components [15]. The generation process is summarized in the algorithm in Fig. 2.

Deterministic sampling can also be used in place of stochastic sampling by deactivating the randomness source, i.e. removing the Wiener process $\bar{\mathbf{w}}_\tau$ in (3), and scaling the diffusion coefficient g. This turns the reverse SDE into a so-called *probability flow ODE* [8]. A comparison between stochastic and deterministic sampling is presented in Figure 2. We display three realizations of the stochastic and deterministic sampler. Note that because of the stochastic noise injected at each step, two stochastic sampler trajectories starting at the same initial state may end up reaching two different modes of the target data distribution, whereas the corresponding mean trajectory systematically reaches the same mode. This suggests that a stochastic sampler could be used to obtain more diverse

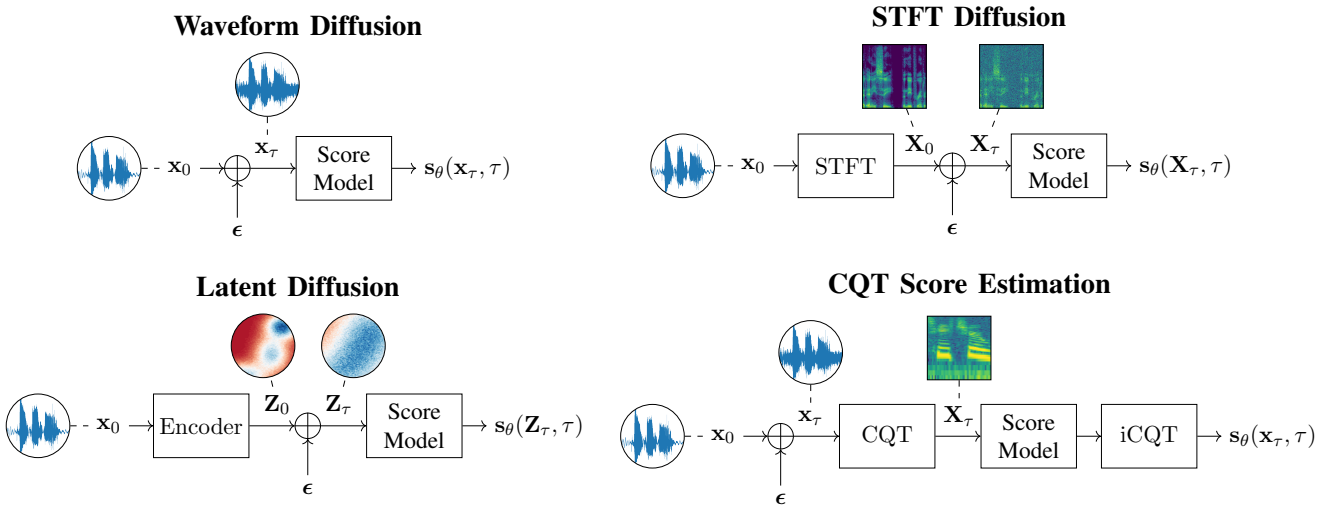


Fig. 3: A diffusion model may be trained in (top-left) waveform [16], (top-right) STFT [17], (bottom-left) latent [21], or (bottom-right) CQT [9] domains. Sampling from the transition kernel $q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)$ can be realized by rescaling the clean data sample $\mathbf{x}_0$ and adding Gaussian noise $\boldsymbol{\epsilon}$ with standard deviation $\sigma(\tau)$ (see (2)). In the top-left, top-right, and bottom-left figures, the noise and the score functions are in the same domain. In the bottom-right figure, the diffusion process is formulated in the time domain but the score model pipeline includes a CQT and its inverse.

samples and also to improve mode coverage, i.e. better represent the modes of the data distribution $p(\mathbf{x}_0)$ regardless of the initial state.

Diffusion processes can be defined for various data representations, depending on the audio application considered. Early works such as [10], [11], [16] directly use the waveform representation, whereas some speech enhancement approaches employ the complex short-time Fourier transform (STFT) domain [12], [17], [18], and several music restoration works consider the Constant-Q Transform (CQT) domain, which is a natural space for harmonic music signals [9], [19], [20]. Learned domains like, e.g., auto-encoder latent spaces, can also be exploited for diffusion to reduce the dimensionality of the original audio data or leverage auto-encoding properties, which gives birth to *latent diffusion models* [21]. Figure 3 offers a schematic overview of diffusion models defined in various domains.

It should be noted that a connection between score-based diffusion models and continuous normalizing flows has been drawn by Lipman et al., creating a new category of so-called *flow matching* models [22]. Flow matching methods generalize Gaussian diffusion models and allow to design more flexible probability paths (based on e.g. optimal transport) between arbitrary terminal distributions. This approach has been used to train a foundational speech model, which can be finetuned to perform restoration tasks such as speech enhancement and separation [23].

Model-based processing with diffusion models

In statistical model-based speech enhancement, each time-frequency bin of the speech and noise spectrograms is often assumed to be mutually independent and to follow a zero-mean complex Gaussian prior distribution [2]. For an additive mixture model, this yields a Gaussian likelihood model for the mixture and a Gaussian posterior model

for the clean speech estimate using Bayes' rule. Under this Gaussian assumption, the posterior mean can be derived as the celebrated Wiener filter solution, providing the optimal speech estimate in the minimum mean square error (MMSE) sense. However, distributional and independence assumptions are merely approximations utilized out of convenience for the derivation of closed-form estimators, e.g. the mentioned Wiener filter. With diffusion models, there are no distributional and independence assumptions on the speech and noise signals themselves. Indeed, the very intent of deep generative modeling is to allow more flexibility by inferring the signal structure from data rather than the parameters of a fixed distribution.

In contrast to other deep learning approaches to audio restoration, two aspects make diffusion models well suited for the introduction of domain knowledge, showcasing them as model-based approaches. The first property is derived from the physical inspiration of diffusion models and their connection to Gaussian denoising [24], which makes them easier to interpret in comparison to other deep generative models such as GANs. In particular, the Gaussian parameterization of the transition kernel $q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)$ enables the injection of knowledge in the form of specific schedules for the mean $\boldsymbol{\mu}$ and standard deviation σ [11], [17], [25]. Furthermore, domain knowledge can be leveraged to posit a distributional hypothesis for the noise process $\mathbf{w}_\tau$ used during forward and reverse diffusion. For instance, Nachmani et al. [26] propose a Gamma distribution instead of the usual Wiener process $\mathbf{w}_\tau$ with Gaussian increments, as it better fits the estimation error distribution. The authors consequently show improvements in speech generation quality compared to the Gaussian case.

The second powerful property of diffusion models is their natural integration within stochastic optimization and posterior sampling using Bayes' theorem, making them particularly suited for conditional generation. We consider the case of audio restoration under the scope of inverse problem solving, i.e. retrieval of clean audio $\mathbf{x}_0$ from a measurement $\mathbf{y}$. There, an approximation of the measurement likelihood $p(\mathbf{y}|\mathbf{x}_0)$ can be obtained via a closed-form model of the operation corrupting $\mathbf{x}_0$ into $\mathbf{y}$. Combining this likelihood model and the learned deep generative prior with Bayes' rule can provide sampling or stochastic optimization algorithms for the conditional generation of samples from the posterior distribution $p(\mathbf{x}_0|\mathbf{y})$ [9], [20], [27], [28].

In summary, first, we see that the data-driven nature of diffusion models allows a higher degree of versatility than traditional signal processing methods, which are often strictly based on simple closed-form distributions and independence assumptions. Secondly, it is important to note that diffusion models transcend the stereotype of being non-interpretable black-boxes. Instead, they benefit from strong integration within stochastics and enable significant potential for the injection of domain knowledge for model-based audio processing.

CONDITIONAL GENERATION WITH DIFFUSION MODELS

One of the most fundamental uses of diffusion models is to perform unsupervised learning from a finite collection of samples to learn an underlying complex data distribution. This provides the ability of *unconditional generation*,

i.e., to generate new samples from the learned data distribution. To solve audio restoration tasks, a diffusion model must be adapted to generate audio that not only conforms to the learned clean audio distribution but, importantly, is also a plausible reconstruction of a given corrupted signal. This effectively requires the diffusion model to perform *conditional generation*. We distinguish between three families of approaches for diffusion-based generative audio restoration: *(i) input conditioning*, where the score model is provided with a task-specific conditioning signal as input, *(ii) task-adapted diffusion*, where the forward and reverse diffusion processes are modified to interpolate between clean and corrupted signals, and *(iii) external conditioning*, where the score model is trained purely on clean audio data and is later combined with an external conditioner during inference. Approaches that use input conditioning (i) or external conditioning (iii) often initialize the iterative generation process with pure Gaussian noise, and then generate a clean signal by iteratively filtering this noise while being guided by the conditioning signal. In contrast, in task-adapted diffusion (ii), the corrupted audio itself is used for initialization and iteratively filtered, making this approach conceptually closer to a denoising procedure.

(i) Input conditioning: Diffusion models that use input conditioning are provided with a task-specific conditioning signal $\mathbf{c}$ (usually some representation of the corrupted signal $\mathbf{y}$) as an additional input during training and inference. To this end, they employ DNNs as score models that are specifically designed to perform feature fusion between the inputs $\mathbf{x}_\tau$ and $\mathbf{c}$. It should be noted that, in most cases, input conditioning approaches require the use of paired data, as the conditioning signal $\mathbf{c}$ and the target data sample $\mathbf{y}$ should be representations of the same data instance, or at least share some semantics. The earliest works to follow this approach include DiffWave [16], which uses mel-spectrograms as conditioning signals for neural vocoding and text-to-speech tasks. While DiffWave focuses on audio generation rather than restoration, the authors of DiffWave also provide preliminary evidence that an unconditional speech diffusion model can perform speech enhancement by using the corrupted audio $\mathbf{y}$ as a starting point of the sampling process even though the diffusion model was only trained to remove Gaussian noise. DiffuSE [29] builds upon DiffWave to solve speech enhancement tasks, using noisy spectral features as conditioning $\mathbf{c}$.

In the worst case, the score model may not use conditioning $\mathbf{c}$ at all, thus inadvertently performing unconditional rather than conditional generation. One possible solution to this is *classifier-free guidance*, where the conditioning signal is randomly set to zero with a fixed probability during training. This results in a single model that can both provide an estimate for the conditional score $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau|\mathbf{c})$ and the unconditional score $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$. At inference, the two estimates can then be weighted at will to trade quality (more weight on conditional score) for variety (more weight on unconditional score). This idea has been used, for instance, by Liu et al. [21] to perform controllable full-band audio synthesis and can also be employed for various audio restoration tasks.

(ii) Task-adapted diffusion: In many restoration tasks such as denoising, dereverberation and separation, the corrupted signal $\mathbf{y}$ and the clean signal $\mathbf{x}_0$ have same dimensionality. This allows one to define what we denote as

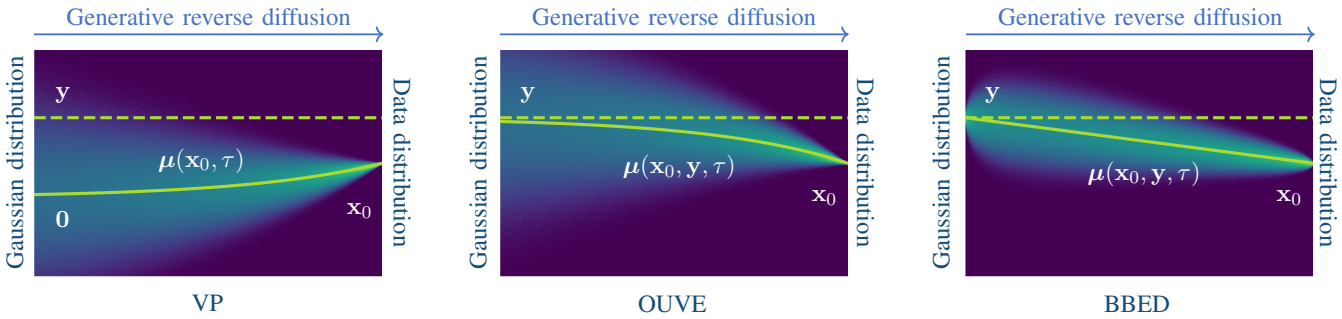


Fig. 4: Comparison of different diffusion processes. (left) Classical variance-preserving (VP) diffusion: mean exponentially interpolates between clean audio $\mathbf{x}_0$ and $\mathbf{0}$, irrespectively of the degraded audio $\mathbf{y}$ [8]. (middle) Task-adapted Ornstein–Uhlenbeck variance exploding (OUVE) diffusion: mean exponentially interpolates between clean audio $\mathbf{x}_0$ and degraded audio $\mathbf{y}$ [12], [17]. (right) Task-adapted brownian bridge with exponential diffusion (BBED): mean linearly interpolates between clean audio $\mathbf{x}_0$ and degraded audio $\mathbf{y}$ [25].

task-adapted diffusion processes, i.e. diffusion processes whose mean $\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, \tau)$ is $\mathbf{x}_0$ at $\tau = 0$ and $\mathbf{y}$ at $\tau = T$, and that interpolates between these terminal values for $\tau \in]0, T[$. This is a form of conditioning which is not introduced as an auxiliary variable to the score model $\mathbf{s}_\theta$ as for input conditioning, but rather directly injected in the parameters of the diffusion process itself. Examples of classical and task-adapted diffusion processes are visualized on Figure 4. CDiffuSE [11] is one of the earliest methods using task-adapted diffusion, formulating the process in discretized time steps. Score-based generative model for speech enhancement (SGMSE) [17] and SGMSE+ [12] extend this idea to the continuous SDE-based formalism of diffusion models to derive pairs of forward and backward processes. Subsequent works [18], [25] build upon this formalism to design alternative forward and backward processes which result in fewer sampling steps and/or higher reconstruction quality. In practice, these methods combine task-adapted diffusion processes with input conditioning, by also providing $\mathbf{y}$ as an auxiliary input to the score model.

The interpolation between $\mathbf{x}_0$ and $\mathbf{y}$ underlying these approaches assumes an additive signal model typical for denoising tasks, which can also be treated as natural for separation tasks [30] or for convolutive corruptions such as reverberation. The aforementioned methods also achieve excellent reconstruction quality for non-additive corruptions like in bandwidth extension [18] and STFT phase retrieval [31], which shows their ability to perform *blind* restoration, i.e. when the corruption operator is not perfectly known during inference.

(iii) External conditioning: External conditioning approaches combine an unconditional diffusion model with an external conditioner that provides a conditioning signal during inference. Since the diffusion model is unconditional, no knowledge of the restoration task is accessed during the training stage and no supervision nor paired data is required. Instead, the task-specific information is injected only at inference by the external conditioner. Therefore, external condition methods can leverage diffusion-based foundation models pre-trained on large-scale data, and adapt them for inference without further re-training. One such type of external conditioner is a pre-trained classifier enabling the combined model to perform class-conditional data generation. For audio restoration, the external

conditioner usually takes the form of a task-specific closed-form measurement model. This results in an overall model that combines a strong data-driven prior for clean audio (score model) with a model-based formulation of the specific restoration task (measurement model). This approach shows good results even when the observation $\mathbf{y}$ is affected by measurement noise [9], [28] and has the advantage of not requiring retraining of the diffusion model for new restoration tasks. These approaches can be applied to blind restoration tasks if a good parameterization of the measurement operator is found. The parameterization enables classical estimation algorithms to be utilized for joint inference of the measurement model and target audio sample estimation, as Moliner et al. [20] accomplished in the blind bandwidth extension of historical music recordings.

Diffusion Models for Inverse Problems

We have seen different strategies to condition diffusion models for audio restoration tasks. This section delves into the *external conditioning* approach, specifically focusing on the application of diffusion models for solving inverse problems in the audio domain. Several audio restoration tasks can be formulated as an inverse problem, wherein an observed audio signal $\mathbf{y}$ is the result of corrupting a clean signal $\mathbf{x}_0$ with a degradation model $\mathcal{A}(\cdot)$ and additive noise $\mathbf{n}$, which can be expressed as

$$\mathbf{y} = \mathcal{A}(\mathbf{x}_0) + \mathbf{n}. \tag{6}$$

This model covers an infinite set of possible degradations, depending on how the operator $\mathcal{A}(\cdot)$ is defined. Three cases of particular interest are showcased in Figure 5. Initially, we concentrate on scenarios in which both the degradation model $\mathcal{A}(\cdot)$ and the noise statistics $\mathbf{n}$ are known. The goal is to recover the original signal $\mathbf{x}_0$ from the corrupted observations $\mathbf{y}$. However, in many cases, the problem is ill-posed, lacking a unique solution and defying straightforward resolution.

Often, solving an inverse problem is approached with a maximum a posteriori (MAP) objective

$$\arg\max_{\mathbf{x}_0} p(\mathbf{x}_0|\mathbf{y}) \,, \tag{7}$$

where the posterior distribution factorizes into likelihood and prior $p(\mathbf{x}_0|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}_0)p(\mathbf{x}_0)$. Under a zero-mean Gaussian measurement noise assumption, denoted as $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_y^2\mathbf{I})$, the MAP estimate takes the form

$$\arg\min_{\mathbf{x}_0} \frac{1}{\sigma_y^2}\|\mathbf{y} - \mathcal{A}(\mathbf{x}_0)\|_2^2 + \mathcal{R}(\mathbf{x}_0) \,, \tag{8}$$

where the first term is a reconstruction cost function, in this case an L^2-norm, designed to preserve fidelity with the observations $\mathbf{y}$. The second term, $\mathcal{R}(\mathbf{x}_0)$, functions as a regularizer, incorporating prior information or domain knowledge about the signal. Its purpose is to mitigate the under-determination of the problem by constraining the

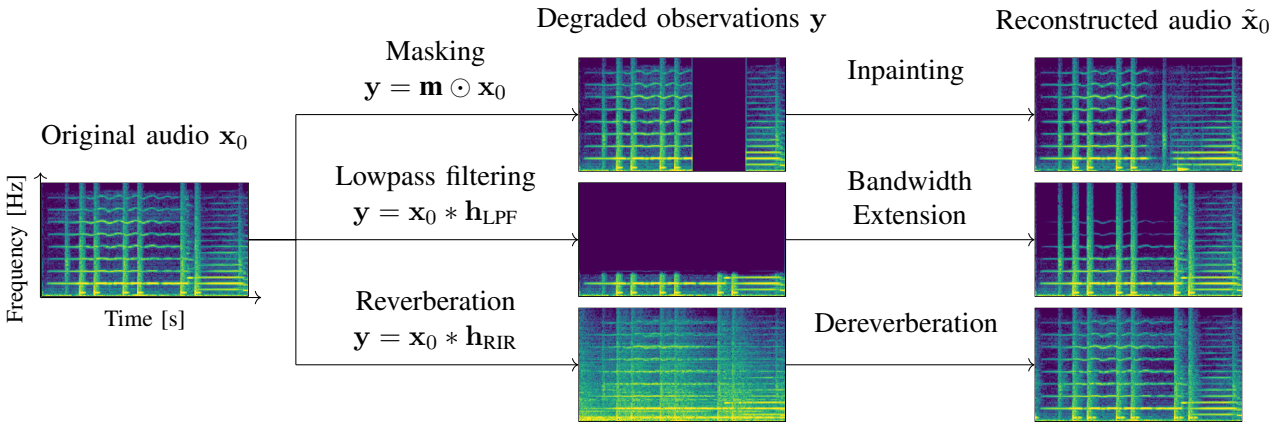


Fig. 5: Visual representation of several inverse problems in audio: (top to bottom) inpainting, bandwidth extension, and dereverberation. (Left) The spectrogram of the original audio signal, $\mathbf{x}_0$, undergoes various transformations via different measurement operators, (middle) the resulting degraded observations, $\mathbf{y}$, correspond to specific audio distortions, and (right) the reconstructed audio signal $\tilde{\mathbf{x}}_0$ is obtained by solving each inverse problem. Notably, the reconstructed example spectrograms (right) closely mirror the original (left), but minor differences appear because of the inherent ill-posed nature of these inverse problems.

space of suitable solutions, thereby making the optimization feasible in practice. In audio processing, a frequently employed regularizer is the sparsity-promoting L^1-norm, which assumes that the true signal is sparse in a specified transform domain, such as time-frequency representations.

A diffusion model learns the statistical characteristics of the training data, in our case of clean audio signals. One can then expect diffusion models to have the potential to serve as strong data-driven priors for solving inverse problems. We will now elaborate on how to leverage these diffusion-based generative priors for solving (8).

To solve an inverse problem using a diffusion model, the score $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$ in the reverse SDE (3) is replaced with the score of the posterior using Bayes' rule

$$\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau|\mathbf{y}) = \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau) + \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{y}|\mathbf{x}_\tau)\,, \tag{9}$$

where the *prior score* $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$ is approximated with the unconditional score model $\mathbf{s}_\theta$ (see (5)). The term $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{y}|\mathbf{x}_\tau)$ represents the *likelihood score*. However, it is important to note that the likelihood $p(\mathbf{y}|\mathbf{x}_\tau)$ is only analytically tractable for $\tau = 0$, as $\mathbf{x}_\tau$ refers to a noisy version of $\mathbf{x}_0$ and the true likelihood is defined through an intractable integral over all possible $\mathbf{x}_0$

$$p_\tau(\mathbf{y}|\mathbf{x}_\tau) = \int_{\mathbf{x}_0} p(\mathbf{y}|\mathbf{x}_0) p_\tau(\mathbf{x}_0|\mathbf{x}_\tau)\, \mathrm{d}\mathbf{x}_0\,. \tag{10}$$

Some works alleviate this issue by simply bypassing the likelihood term, and instead project the state $\mathbf{x}_\tau$ onto the set of the observations $\mathbf{y}$ at every step of the discretized inference process. [8]. The objective of such *projection-based* method is to inject the reliable parts of the observations into the intermediate predictions. This ensures that at each

step, the intermediate output of the algorithm is consistent with the algorithm input, i.e. the degraded audio, which is often refered to as *data consistency* and helps avoiding degenerate solutions. Projection-based methods offer the advantage of ensuring data consistency and simplicity in terms of algorithmic implementation. However, their applicability is limited to a reduced set of linear inverse problems, such as audio inpainting or bandwidth extension [9], [21], where closed-form expressions for the projection step are available.

Other works adopt more theoretically grounded approximations of the likelihood that allow a broader versatility by incorporating a model-based approach. In particular, Chung et al. [27] proposed Diffusion Posterior Sampling, and approximate the likelihood as $p_\tau(\mathbf{y}|\mathbf{x}_\tau) \approx p(\mathbf{y}|\hat{\mathbf{x}}_0(\mathbf{x}_\tau))$. There, $\hat{\mathbf{x}}_0(\mathbf{x}_\tau)$ is a coarse estimate of $\mathbf{x}_0$ obtained by denoising from state $\mathbf{x}_\tau$ in just one deterministic reverse diffusion step. When modeling the measurement noise $\mathbf{n}$ in (6) as a Gaussian $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_y^2\mathbf{I})$, the resulting approximated likelihood is a Gaussian distribution $p\left(\mathbf{y}|\hat{\mathbf{x}}_0(\mathbf{x}_\tau)\right) = \mathcal{N}(\mathcal{A}(\hat{\mathbf{x}}_0(\mathbf{x}_\tau)), \sigma_y^2\mathbf{I})$. It follows that the likelihood score can be computed as:

$$\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{y}|\mathbf{x}_\tau) \approx -\frac{1}{\sigma_y^2}\nabla_{\mathbf{x}_\tau} \|\mathbf{y} - \mathcal{A}\left(\hat{\mathbf{x}}_0(\mathbf{x}_\tau)\right)\|_2^2 . \tag{11}$$

The L^2-norm in (11) can be replaced by any other objective function that better fits the statistics of the measurements [27]. For example in [32], the measurement noise $\mathbf{n}$ is modelled as a Gaussian in the compressed STFT domain, where the compression is a square-root power law on the STFT magnitude, leaving the phase unchanged. This helps accounting for the heavy-tailedness of speech distributions [33]. It is important to note that the gradient operator $\nabla_{\mathbf{x}_\tau}$ requires differentiating through the score model by backpropagation, which introduces a computational overhead. In practice, the unknown measurement noise variance σ_y^2 is estimated empirically using e.g. the norm of the gradients in (11) [9]. Compared to projection-based methods, this approach is not limited to linear problems and can be applied to cases where $\mathcal{A}(\cdot)$ is nonlinear, as long as the operator $\mathcal{A}(\cdot)$ is differentiable. A geometrical perspective on the sampling process is displayed in Figure 6. This diagram illustrates the intuition behind conditional sampling with a diffusion model, in this case in the context of bandwidth extension. This strategy has been successfully applied in audio bandwidth extension [9], audio inpainting [19], and dereverberation [28].

Blind inverse problems

Until this point, our analysis has proceeded under the assumption that the degradation operator $\mathcal{A}(\cdot)$ is known. However, in practical applications, the degradation operator is often unknown. This lack of knowledge about the degradation operator renders the calculation of the posterior $p(\mathbf{x}_0|\mathbf{y})$ a *blind* inverse problem, substantially raising the difficulty of the task. The Diffusion Posterior Sampling approach [27], as previously explained, provides a valuable foundation that can be extended to tackle blind inverse problems. In scenarios where we possess at least some knowledge of the structure of the degradation operator, a viable strategy is to embrace a model-based approach.

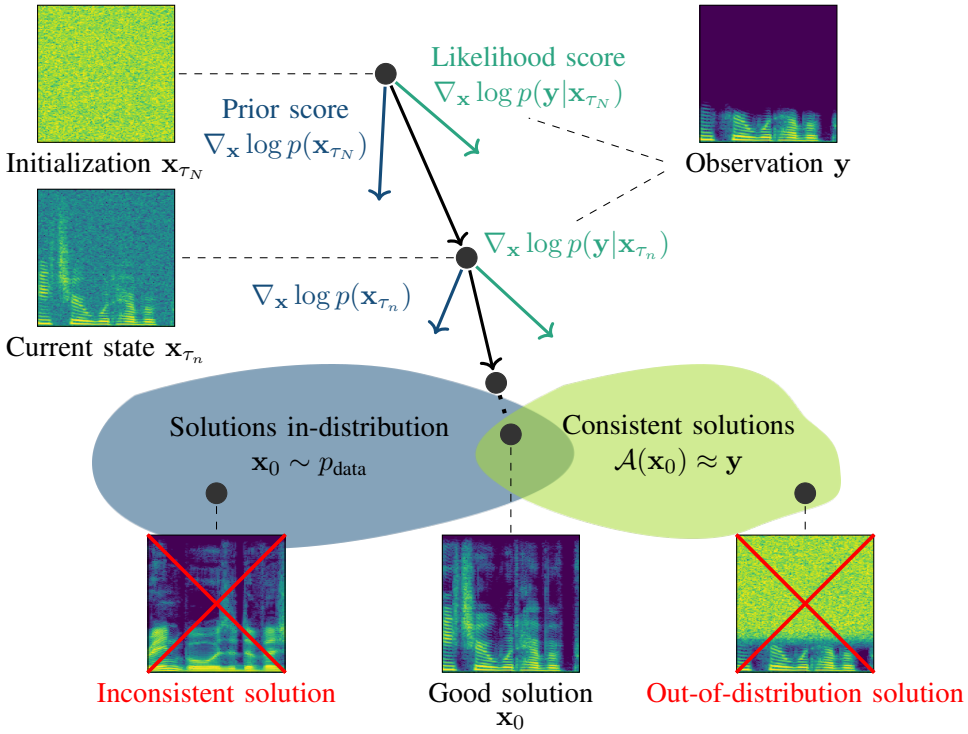


Fig. 6: Geometrical interpretation of posterior sampling with diffusion models (e.g. [27]). The prior score guides the trajectories towards solutions within the training data manifold, or in-distribution with the training data (gray space). Simultaneously, the role of the likelihood score is to steer the sampling trajectories toward a solution space consistent with the observed data (light green space). When properly weighted, the two components pull the sampling process to the intersection of these two manifolds. This intersection exists and contains the solutions to the inverse problem if the two score functions are properly estimated and if the solutions are contained in the manifold spanned by the training data, i.e. if the training dataset is properly adapted to the problem.

This involves designing a parametric model of the degradation operator, denoted as $\mathcal{A}_\phi(\cdot)$, and jointly optimizing its parameters ϕ alongside the restored audio signal throughout the sampling process.

An example of this approach is the Blind Audio Bandwidth Extension (BABE) [20], which addresses the problem of blind reconstruction of missing high frequencies in music from bandlimited observations without knowledge of the lowpass degradation, such as the cutoff frequency. This challenge is typical in restoring historical audio recordings. In BABE, the measurement model $\mathcal{A}_\phi(\cdot)$ is parameterized by a piecewise approximation of a low-pass filter in the frequency domain, where the parameters ϕ represent the cutoff frequencies and decay slopes of this filter [20]. The optimization process, as illustrated in Figure 7, alternates between sampling updates of the audio signal $\mathbf{x}_\tau$ and refining ϕ through stochastic gradient descent, using a maximum likelihood objective as the guiding principle. BUDDy [32] takes a similar approach and solves joint speech dereverberation and room acoustics estimation by combining Diffusion Posterior Sampling with a model-based subband filter approximating room impulse response. The resulting method largely outperforms other blind unsupervised dereverberation methods. Thanks to unsupervised learning, BUDDy seamlessly adapts to new acoustic scenarios, whereas supervised methods typically struggle when

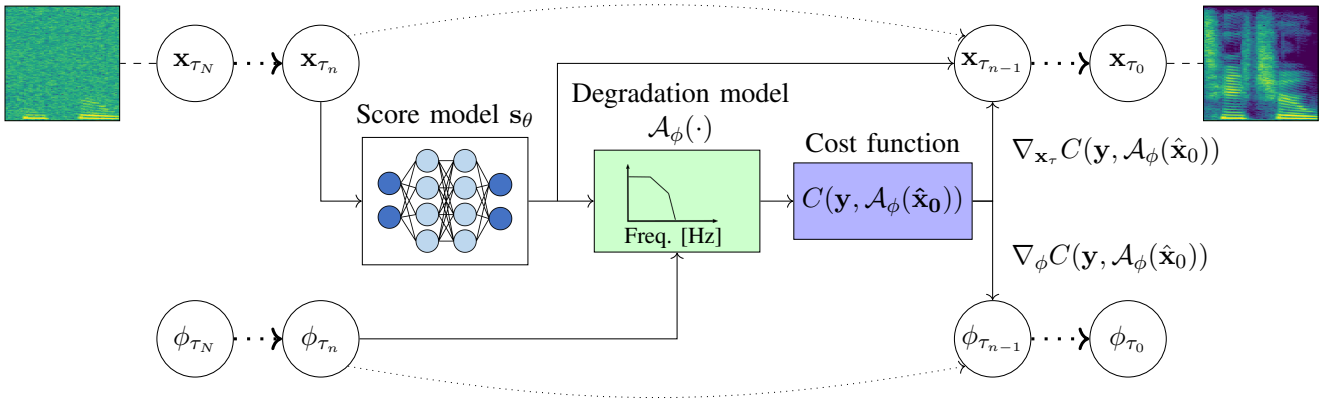


Fig. 7: BABE: posterior sampling algorithm for solving blind bandwidth extension using a prior score model $\mathbf{s}_\theta$ and a parameterized degradation operator $\mathcal{A}_\phi$ [20]. The optimization alternates between updating the reconstructed signal $\mathbf{x}$ (top) and the degradation parameters ϕ (bottom). For ease of reading, we write $\hat{\mathbf{x}}_0 := \hat{\mathbf{x}}_0(\mathbf{x}_{\tau_n})$.

there is a mismatch between training and testing conditions.

Practical Requirements of Diffusion-based Sampling for Audio Tasks

While diffusion models provide powerful priors that can be employed for various audio restoration tasks, they require some improvements to be suitable for real-time acoustic communications. We divide these requirements into two categories: *(i) inference speed and causal processing*, which can be prohibitive for low-latency real-time applications, and *(ii) robustness to adverse conditions*, which must be assured for integration into reliable systems.

Inference speed and causal processing

One major drawback of diffusion models is their slow inference. As the score model is called at each step of the reverse process, the computational complexity is directly proportional to the number of steps used and the order of the solver, i.e., the number of score estimations used per time step. Using more diffusion steps naturally provides better sample reconstruction since the truncation error of the numerical solver is reduced when the step size is decreased. Similarly, increasing the solver order reduces the per-step truncation error. However, both these options lead to an increased computational cost. Furthermore, accumulating truncation errors over the diffusion trajectory can make the samples diverge from the distribution learned during training, and therefore make the score model produce unreliable estimates, which is referred to as the *drifting bias*. These two sources of error compound over the diffusion trajectory, therefore, without further optimization, high-quality reconstruction can only be obtained at a high computational cost. This section presents several methods to reduce the computational complexity of diffusion-based methods in audio applications.

Reducing per-step inference time: A natural way to accelerate inference is to reduce the cost of each call to the score model. This can be obtained by minimizing the size of the neural network used for score inference through e.g.

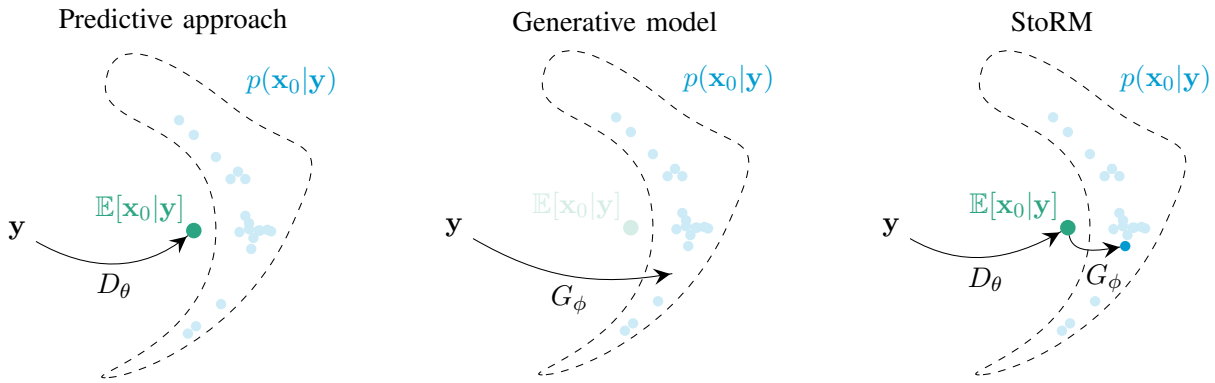


Fig. 8: Visualization of the inference process for the predictive, generative and StoRM [18] models for a complex posterior distribution. With the proposed two-stage inference, StoRM uses the predictive mapping to the posterior mean $\mathbb{E}[\mathbf{x}_0|\mathbf{y}]$ as an intermediate step for generation of a sample which is more likely to lie in high-density regions of the posterior $p(\mathbf{x}_0|\mathbf{y})$.

knowledge distillation, or by reducing the size of the space itself where diffusion is performed, resulting in *latent diffusion models*. The latent space should be designed such that its reduced dimensionality has a limited impact on the reconstruction quality, and its structure allows for score estimation with a reasonably-sized neural network. Latent diffusion is popular in text-to-audio generation and has been recently applied to audio editing (including restoration) in AUDIT [34], which uses latents provided by a VAE.

Improving initialization: Another possibility to accelerate sampling is to find a better initial prediction to reduce the distance between the initial condition $\mathbf{x}_T$ and the target sample $\mathbf{x}_0$. This can be provided by a separate plug-in predictive network providing an estimate of the posterior mean $\mathbb{E}[\mathbf{x}_0|\mathbf{y}]$ as proposed by Lemercier et al. in their Stochastic Regeneration Model (StoRM) [18] for speech enhancement (see Figure 8). The diffusion-based generative model can restore target cues potentially destroyed by the predictive stage while additionally removing residual corruption. The resulting approach requires significantly fewer function evaluations than the original diffusion-only model in [12], for a better-sounding result. Figure 9 shows the clean, degraded, and restored speech spectrograms produced with StoRM. As a simpler alternative, the corrupted utterance $\mathbf{y}$ can be directly used as the mean of the initial state $\mathbf{x}_T$. This latter strategy is sometimes referred to as *warm initialization* and has already been used in audio-related tasks such as speech enhancement [11] and bandwidth extension [20]. A good initial prediction can also be obtained by designing a more suitable diffusion trajectory to reduce the mismatch between training and inference, as suggested by Lay et al. [25] for speech enhancement. As shown in Figure 4, the BBED diffusion process proposed in [25] has a linear, constant speed mean interpolating between the clean and noisy speech, which effectively terminates at the clean speech in finite time, unlike the original OUVE diffusion process proposed in [17].

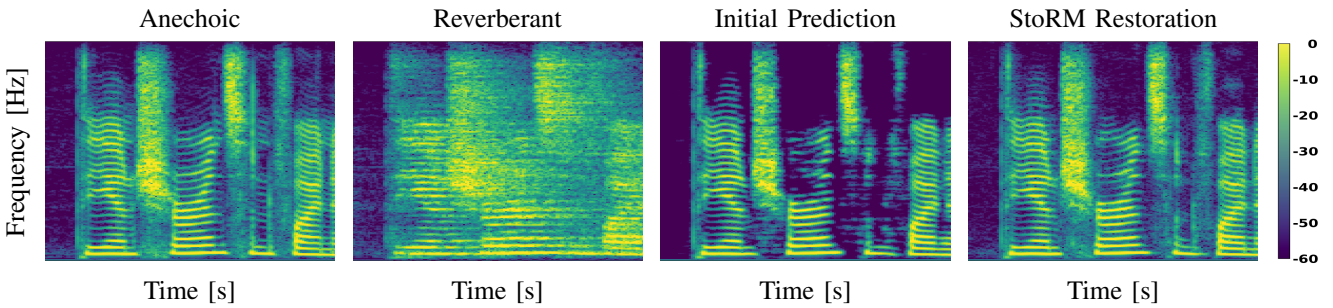


Fig. 9: Dereverberation results with StoRM [18]. Input T_{60} is 1.06 s. Three seconds of audio are shown, and the bandwidth is 8 kHz. Severe speech distortions are observed in the initial prediction because of the harsh reverberant conditions. StoRM corrects the distortions and restores the formant structure without residual reverberation.

Reducing the number of steps: The remaining approaches investigate how to reduce the number of diffusion steps of the reverse process. As in most ODE/SDE integration problems, using off-the-shelf higher-order samplers can improve the per-step precision but here it comes at the cost of more calls to the neural network for each step, which leads to a non-trivial tradeoff between computational complexity and sample quality. In denoising diffusion implicit models (DDIM) [35] instead, the Markovian property of the transition kernel is deliberately removed by conditioning the next reverse diffusion estimate $\mathbf{x}_{\tau_{n-1}}$ on both the previous state $\mathbf{x}_{\tau_n}$ and $\hat{\mathbf{x}}_0$, a coarse estimate of the clean signal obtained via one-step denoising (see the section above on inverse problems). This allows an arbitrary number of steps to be skipped during reverse diffusion, which can significantly accelerate inference.

A progressive distillation method for reverse diffusion is used for text-to-speech generation in [36]. Leveraging DDIM sampling, a new student diffusion sampler learns at each iteration of the distillation process how to perform reverse diffusion using half as many steps as the current teacher. The resulting distilled sampler generates speech with similar quality as the original sampler using 64x more steps.

The noise variance schedule and time discretization used for reversed diffusion can also be optimized to reduce the number of steps, instead of being pre-defined. In [37], the schedule is learned by training an auxiliary hyper-network on top of existing denoising diffusion models. The resulting approach enables impressive speech generation results in as few as three reverse diffusion steps.

Finally, some auxiliary losses and training schemes are designed to ensure that the diffusion states remain as close as possible to the domain seen by the score network during training, thereby mitigating the so-called drifting bias. Lay et al. [38] propose a two-stage training method for diffusion-based speech enhancement following such a concept. The score network is first trained with denoising score matching and then fine-tuned to overfit a particular reverse diffusion sampler, by matching the final estimate of the solver to the clean speech target. High-quality speech enhancement is obtained with as few as one reverse diffusion step, reaching real-time computational complexity.

Causal processing: In real-time acoustic communications (e.g. hearing aids), future information can not be

used to process the current signal which means processing must be causal. Diffusion models can be adapted for causal processing, as in Richter et al. [39], where the convolutional score network architecture and the audio level normalization procedure are modified to meet causality requirements.

Robustness to adverse conditions

Artifacts produced by diffusion models can differ in nature from those produced by statistical signal processing methods or predictive deep learning models. It was observed in [12] that speech enhancement diffusion models tend to hallucinate for negative input signal-to-noise ratios, i.e. when noise dominates clean speech. This can lead to speech inpainting in noise-only regions, breathing and gasping artifacts, or the introduction of phonetic confusion, which may have a negative impact in real-world applications. This behavior can be mitigated by introducing external modalities such as video in Richter et al. [40], where lip movements are analyzed to determine the phoneme used as conditioning for score estimation guidance. Alternatively, as presented in StoRM [18] the input signal-to-noise ratio can be first increased by using a predictive deep learning model to remove parts of the noise, at the potential cost of speech distortions. A generative diffusion model is then used to reconstruct the noisy and distorted speech, which was shown to help avoid hallucination effects and thus increase the robustness to challenging conditions.

Generative pre-training is another approach to increase robustness to outliers. It involves using a pretext task such as masked modeling to train the diffusion model in a self-supervised fashion. Masked modeling involves randomly masking some regions of audio and instructing the model to fill in those masked sections using the available context information, i.e. the non-masked regions. This pre-trained model can then be fine-tuned for a particular downstream task (e.g. speech enhancement, music restoration, etc.) using a supervised setting. Liu et al. [23] show that their diffusion model SpeechFlow benefits from generative pre-training, as it increases its robustness to adverse scenarios such as noise-dominated utterances in speech enhancement. They also notice that generative pre-training consistently increases performance for most speech restoration tasks.

Finally, running several realizations of the reverse diffusion process and measuring the empirical standard deviation of the obtained estimates can provide the user with a natural measure of uncertainty, which can help detect outliers and estimate the robustness of the approach on the given task.

CONCLUSION

This article discussed diffusion models as deep conditional generative models for audio restoration. We suggested that diffusion models can be considered as serious candidates for model-based audio processing, as we recalled that domain knowledge can be injected into various aspects of their design such as parameterization of diffusion trajectories, or modeling of a measurement likelihood for posterior sampling with diffusion priors. By categorizing the various forms of conditioning proposed in diffusion approaches—namely input conditioning, task-adapted

processes, and external conditioning—we highlight the structural flexibility of diffusion models and their resulting appreciable degree of interpretation. In particular, looking at audio restoration under the scope of solving inverse problems, we showed that we can combine diffusion models with Bayesian tools and stochastic optimization, thereby leveraging various parameterizations of degradation operators for informed and blind inverse problems. The quality of diffusion-based audio generation is remarkable, and although this can be originally outbalanced by disadvantages regarding practical requirements, e.g., robustness to adverse conditions or inference speed, we exposed several approaches and studies solving these drawbacks. We believe these solutions can be combined to yield robust, fast diffusion models for real-time acoustic communications.

ACKNOWLEDGMENTS

This work has been funded by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169), DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the Grant-No. HIDSS-0002, and NordicSMC (Nordic Sound and Music Computing Network) with NordForsk project 86892.

REFERENCES

[1] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration—A Statistical Model Based Approach*. Springer, 1998.

[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.), John Wiley & Sons, 2018.

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

[4] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Proc. Neural Inf. Process. Syst.*, vol. 27, pp. 139–144, 2014.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *Proc. Int. Conf. Learning Repr.*, pp. 1–14, 2014.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Neural Inf. Process. Syst.*, pp. 6840–6851, 2020.

[8] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learning Repr.*, pp. 1–36, 2021.

[9] E. Moliner, J. Lehtinen, and V. Välimäki, "Solving audio inverse problems with a diffusion model," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1–5, 2023.

[10] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," *Proc. Int. Conf. Learning Repr.*, pp. 1–15, 2021.

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

[12] 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 Lang. Process.*, vol. 31, pp. 2351–2364, 2023.

[13] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*. Springer, 2013.

[14] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[15] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 3rd ed., 2007.

[16] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," *Proc. Int. Conf. Learning Repr.*, pp. 1–17, 2021.

[17] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Proc. Interspeech*, 2022.

[18] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 31, pp. 2724–2737, 2023.

[19] E. Moliner and V. Välimäki, "Diffusion-based audio inpainting," *J. Audio Eng. Soc.*, vol. 72, pp. 100–113, Mar. 2024.

[20] E. Moliner, F. Elvander, and V. Välimäki, "Blind audio bandwidth extension: A diffusion-based zero-shot approach," *arXiv:2306.01433*, 2024.

[21] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proc. Int. Conf. Machine Learning*, pp. 21450–21474, 2023.

[22] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *Proc. Int. Conf. Learning Repr.*, pp. 1–28, 2023.

[23] A. H. Liu, M. Le, A. Vyas, B. Shi, A. Tjandra, and W.-N. Hsu, "Generative pre-training for speech with flow matching," in *Proc. Int. Conf. Learning Repr.*, pp. 1–20, 2024.

[24] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. Neural Inf. Process. Syst.*, pp. 1–13, 2019.

[25] B. Lay, S. Welker, J. Richter, and T. Gerkmann, "Reducing the prior mismatch of stochastic differential equations for diffusion-based speech enhancement," in *Proc. Interspeech*, pp. 1–5, 2023.

[26] E. Nachmani, R. S. Roman, and L. Wolf, "Denoising diffusion gamma models," in *Proc. Int. Conf. Learning Repr.*, pp. 1–13, 2022.

[27] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," in *Proc. Int. Conf. Learning Repr.*, pp. 1–30, 2023.

[28] J.-M. Lemercier, S. Welker, and T. Gerkmann, "Diffusion posterior sampling for informed single-channel dereverberation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pp. 1–5, 2023.

[29] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *Proc. Asia-Pacific Signal and Information Processing Association (APSIPA)*, pp. 659–666, 2021.

[30] R. Scheibler, Y. Ji, S.-W. Chung, J. Byun, S. Choe, and M.-S. Choi, "Diffusion-based generative speech source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023.

[31] T. Peer, S. Welker, and T. Gerkmann, "DiffPhase: Generative diffusion-based STFT phase retrieval," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1–5, IEEE, 2023.

[32] E. Moliner, J.-M. Lemercier, S. Welker, T. Gerkmann, and V. Välimäki, "BUDDy: Single-channel blind unsupervised dereverberation with diffusion models," in *Proc. Int. Workshop Acoustic Signal Enhancement*, pp. 1–5, 2024.

[33] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proc. Int. Workshop Acoustic Signal Enhancement*, pp. 1–4, 2010.

[34] Y. Wang, Z. Ju, X. Tan, L. He, Z. Wu, J. Bian, and S. Zhao, "Audit: Audio editing by following instructions with latent diffusion models," in *Proc. Neural Inf. Process. Syst.*, pp. 71340–71357, 2023.

[35] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learning Repr.*, pp. 1–22, 2022.

[36] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, and Y. Ren, "ProDiff: Progressive fast diffusion model for high-quality text-to-speech," in *ACM Multimedia*, 2022.

[37] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "BDDM: Bilateral denoising diffusion models for fast and high-quality speech synthesis," in *Proc. Int. Conf. Learning Repr.*, pp. 1–22, 2022.

[38] B. Lay, J.-M. Lemercier, J. Richter, and T. Gerkmann, "Single and few-step diffusion for generative speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 636–630, 2024.

[39] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Causal diffusion models for generalized speech enhancement," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 780–789, 2024.

[40] J. Richter, S. Frintrop, and T. Gerkmann, "Audio-visual speech enhancement with score-based generative models," in *Proc. ITG Conf. Speech Communication*, pp. 275–279, 2023.

4.2 Analyzing Predictive Approaches versus Diffusion-based Generative Models for Speech Restoration [P7]

Abstract

Diffusion-based generative models have had a high impact on the computer vision and speech processing communities these past years. Besides data generation tasks, they have also been employed for data restoration tasks like speech enhancement and dereverberation. While discriminative models have traditionally been argued to be more powerful e.g. for speech enhancement, generative diffusion approaches have recently been shown to narrow this performance gap considerably. In this paper, we systematically compare the performance of generative diffusion models and discriminative approaches on different speech restoration tasks. For this, we extend our prior contributions on diffusion-based speech enhancement in the complex time-frequency domain to the task of bandwith extension. We then compare it to a discriminatively trained neural network with the same network architecture on three restoration tasks, namely speech denoising, dereverberation and bandwidth extension. We observe that the generative approach performs globally better than its discriminative counterpart on all tasks, with the strongest benefit for non-additive distortion models, like in dereverberation and bandwidth extension. Code and audio examples can be found online.

Reference

Jean-Marie Lemercier, Julius Richter, Simon Welker and Timo Gerkmann, "Analysing Diffusion-based Generative Approaches versus Discriminative Approaches for Speech Restoration", *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Rhodes Island, Greece, 2023, DOI: 10.1109/ICASSP49357.2023.10095258

Copyright Notice

Authors' Contributions

Jean-Marie Lemercier is the first author of this publication. He implemented all algorithms, trained the neural networks used in the paper, conducted the experimental validation, and wrote the manuscript. Julius Richter and Simon Welker brought their feedback on all methods through discussions, they also helped with reviewing the manuscript, and are at the origin of the conditional diffusion model SGMSE+ used in the experiments. Timo Gerkmann brought insights on the experimental validation, mathematical derivations, and reviewed the manuscript.

ANALYSING DIFFUSION-BASED GENERATIVE APPROACHES VERSUS DISCRIMINATIVE APPROACHES FOR SPEECH RESTORATION

Jean-Marie Lemercier[⋆]*, Julius Richter*[⋆]*, Simon Welker*[⋆×]*, Timo Gerkmann*[⋆]

[⋆]Signal Processing (SP), Universität Hamburg, Germany
[×] Center for Free-Electron Laser Science, DESY, Germany
{firstname.lastname}@uni-hamburg.de

ABSTRACT

Diffusion-based generative models have had a high impact on the computer vision and speech processing communities these past years. Besides data generation tasks, they have also been employed for data restoration tasks like speech enhancement and dereverberation. While discriminative models have traditionally been argued to be more powerful e.g. for speech enhancement, generative diffusion approaches have recently been shown to narrow this performance gap considerably. In this paper, we systematically compare the performance of generative diffusion models and discriminative approaches on different speech restoration tasks. For this, we extend our prior contributions on diffusion-based speech enhancement in the complex time-frequency domain to the task of bandwith extension. We then compare it to a discriminatively trained neural network with the same network architecture on three restoration tasks, namely speech denoising, dereverberation and bandwidth extension. We observe that the generative approach performs globally better than its discriminative counterpart on all tasks, with the strongest benefit for non-additive distortion models, like in dereverberation and bandwidth extension. Code and audio examples can be found online[1].

Index Terms— generative modelling, diffusion models, speech enhancement, dereverberation, bandwidth extension

1. INTRODUCTION

Speech corruptions arise in real-life scenarios and modern communication devices, when clean speech sources are impacted by background noise, interfering speakers, room acoustics and channel degradation. Speech restoration therefore aims at recovering clean speech from the corrupted signal. Traditional speech restoration methods leverage the different statistical properties of the target and interference signals [1]. Data-driven approaches based on machine learning predominately employ discriminative models that learn a single best deterministic mapping between corrupted speech and the corresponding clean speech target [2].

In contrast, generative models implicitly or explicitly learn the target distribution and allow to generate multiple valid estimates instead of a single best estimate as in discriminative approaches [3]. For example, diffusion-based generative models, or simply *diffusion models*, have shown great success in learning the data distribution of natural images [4, 5, 6]. This class of models uses a *forward process* to slowly turn data into a tractable prior, such as a standard normal distribution, and train a neural network to solve the *reverse process* to generate clean data from this prior. These diffusion models can also be used for conditional generation in restoration tasks, which has recently been proposed for speech enhancement and dereverberation [7, 8, 9, 10]. They can in that regard be functionally seen as a mean of generating clean speech based on noisy speech, and can be thus compared to discriminative approaches. However, to make a fair comparison of these two conceptually different approaches, similar network architectures and same training data should be used.

In this work, we present an analysis of a generative diffusion model as compared to its discriminative counterpart sharing the same deep neural network (DNN) architecture, for various speech restoration tasks. We use our previous method which defines the diffusion process in the complex spectrogram domain [7, 8]. We show that the performance gap between the generative and discriminative models varies with respect to the corruption at hand. We evaluate our proposed approaches on the WSJ0 corpus, using various simulated corruptions and recorded background noise. Finally we compare our bandwidth extension model with state-of-the-art bandwidth extension methods on the VCTK corpus.

The remainder of the paper is organized as follows. We first present the three speech restoration problems benchmarked, along with popular solutions for solving them. Then, we introduce diffusion-based generative models using the stochastic differential equation (SDE) formalism. We continue by explaining our experimental setup including data generation and training methods. Finally, we present and discuss our results.

2. SPEECH RESTORATION TASKS AND RELATED WORK

2.1. Speech enhancement

Speech enhancement consists in removing an additive interference n (e.g. background noise or interfering speakers) from the corrupted mixture y to extract the clean speech target s:

$$y = s + n \tag{1}$$

Popular enhancement methods include Wiener-inspired spectral filtering [1], discriminative machine learning methods [2] or generative approaches like denoising variational auto-encoders (VAEs) [11]. Recently, diffusion models were proposed to tackle speech enhancement either in the time domain [12] or in the complex time-frequency (T-F) domain [7, 10, 8].

2.2. Speech dereverberation

Reverberation is caused by room acoustics, and is characterized by multiple reflections on the room enclosures. Late reflections particularly degrade the speech signal and may result in reduced intelligibility [13].

This work has been funded by the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380, DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the Grant-No. HIDSS-0002, and the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169).

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

The corruption model is then convolutive, as the clean speech s is convolved with a room impulse response (RIR) h representing the acoustic path between the source and the listener:

$$y = s * h \tag{2}$$

Single-channel dereverberation methods range from spectral enhancement [14], inverse filtering [15], and cepstral processing [16] to machine learning algorithms using DNNs in the complex T-F domain [17] or in the time-domain [18].

2.3. Bandwidth extension

Audio super-resolution, or bandwidth extension, aims at converting a low-sampling rate signal back to a version sampled at a higher rate, regenerating time resolution, high-frequency content and audio quality. The corruption process is linear and involves an anti-aliasing low-pass filter followed by a decimation operation:

$$y = \mathrm{Resample}(s * a,\ f_s^{\mathrm{up}},\ f_s^{\mathrm{low}}) \tag{3}$$

where a is the anti-aliasing filter impulse response, f_s^{up} the original high sampling rate and f_s^{low} the low sampling rate.

Several discriminative methods were proposed to tackle bandwidth extension for speech signals [19, 20]. Generative approaches based on neural vocoders using generative adversarial networks (GANs) were also proposed [21, 22, 23]. A continuous-time diffusion model in the time-domain was proposed in [24].

3. SCORE-BASED DIFFUSION MODELS FOR SPEECH RESTORATION

Score-based diffusion models are defined by three components: a forward diffusion process, a score estimator and a sampling method for inference.

3.1. Forward and reverse processes

The stochastic forward process $\{\mathbf{x}_t\}_{t=0}^T$ is modeled as the solution to a SDE, in the Itô sense [25, 26]:

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w} \tag{4}$$

where $\mathbf{x}_t$ is the current state of the process indexed by a continuous time variable $t \in [0,T]$ with the initial condition $\mathbf{x}_0$ representing clean speech. As our process is defined in the T-F domain, the variables in bold are assumed to be one-dimensional vectors in $\mathbb{C}^d$ containing the coefficients of a flattened complex spectrogram, whereas variables in regular font represent real scalar values. The stochastic process $\mathbf{w}$ is a standard d-dimensional Brownian motion, which implies that $\mathrm{d}\mathbf{w}$ is a zero-mean Gaussian random variable with standard deviation $\sqrt{\mathrm{d}t}$ for each T-F bin.

The *drift* function $\mathbf{f}$ and *diffusion* coefficient g as well as the initial condition $\mathbf{x_0}$ and the final diffusion time T define uniquely the Itô process $\{\mathbf{x}_t\}_{t=0}^T$. Under some regularity conditions on $\mathbf{f}$,g allowing a unique and smooth solution to the Kolmogorov equations associated to (4), the reverse process $\{\mathbf{x}_t\}_{t=T}^0$ is another diffusion process defined as the solution of a SDE, with the following form [27, 26]:

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, t) + g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\right] \mathrm{d}t + g(t)\,\mathrm{d}\bar{\mathbf{w}}, \tag{5}$$

where $\mathrm{d}\bar{\mathbf{w}}$ is a d-dimensional Brownian motion for the time flowing in reverse and $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ is the *score function*, i.e. the gradient of the logarithm data distribution for the current process state $\mathbf{x}_t$.

Speech restoration tasks can be considered as conditional generation tasks, i.e. generation of clean speech $\mathbf{x_0}$ conditioned by the corrupted speech $\mathbf{y}$. In [7, 8] we proposed to incorporate the conditioning directly into the diffusion process by defining the forward process as the solution to the following Ornstein-Uhlenbeck SDE [25]:

$$\mathrm{d}\mathbf{x}_t = \underbrace{\gamma\,(\mathbf{y} - \mathbf{x}_t)}_{:=\,\mathbf{f}(\mathbf{x}_t,\mathbf{y})}\mathrm{d}t + \underbrace{\left[\sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^t \sqrt{2\log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}\right]}_{:=\,g(t)} \mathrm{d}\mathbf{w}, \tag{6}$$

with γ a *stiffness* hyperparameter, and $\sigma_{\min}$ and $\sigma_{\max}$ two hyperparameters controlling the *noise scheduling*, that is, the amount of Gaussian white noise injected at each timestep of the process.

The interpretation of our forward process in Eq. (6), visualized on Fig. 1, is as follows: at each time step and for each T-F bin independently, an infinitesimal amount of corruption is added to the current process state $\mathbf{x}_t$, along with Gaussian noise with standard deviation $g(t)\sqrt{\mathrm{d}t}$. Given an initial state $\mathbf{x}_0$ and $\mathbf{y}$, the Itô forward process corresponding to the solution of (6) admits a Gaussian distribution for the process state $\mathbf{x}_t$ called *perturbation kernel*:

$$p_{0t}(\mathbf{x}_t|\,\mathbf{x}_0, \mathbf{y}) = \mathcal{N}_{\mathbb{C}}\big(\mathbf{x}_t; \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t), \sigma(t)^2\mathbf{I}\big), \tag{7}$$

where $\mathcal{N}_{\mathbb{C}}$ denotes the circularly-symmetric complex normal distribution and $\mathbf{I}$ the identity matrix. Given the simple Gaussian kernel, closed-form solutions for the mean $\boldsymbol{\mu}$ and variance $\sigma(t)^2$ can be determined [25]:

$$\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t) = \mathrm{e}^{-\gamma t}\mathbf{x}_0 + (1 - \mathrm{e}^{-\gamma t})\mathbf{y}, \tag{8}$$

$$\sigma(t)^2 = \frac{\sigma_{\min}^2\big((\sigma_{\max}/\sigma_{\min})^{2t} - \mathrm{e}^{-2\gamma t}\big)\log(\sigma_{\max}/\sigma_{\min})}{\gamma + \log(\sigma_{\max}/\sigma_{\min})}. \tag{9}$$

3.2. Score function estimator

When performing inference by sampling through the reverse SDE in Eq. (5), the score function $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ is not readily available. Thus, it is approximated by a DNN $\mathbf{s}_\theta$, called the *score model*. In particular, given the Gaussian form of the perturbation kernel $p_{0t}(\mathbf{x}_t|\mathbf{x}_0,\mathbf{y})$ and the regularity conditions exhibited by the mean and variance, a *denoising score matching* objective can be used to train the score model $\mathbf{s}_\theta$ [28].

The score function of the perturbation kernel is:

$$\nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\,\mathbf{x}_0, \mathbf{y}) = -\frac{\mathbf{x}_t - \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t)}{\sigma(t)^2}. \tag{10}$$

Therefore we can reparameterize the denoising score matching objective as follows [26]:

$$\begin{aligned}\mathcal{J}(\theta) &= \mathbb{E}_{t,\mathbf{x}_0,\mathbf{y},\{\mathbf{x}_t|\mathbf{x}_0,\mathbf{y}\}}\left[\left\|\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t) - \nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\,\mathbf{x}_0, \mathbf{y})\right\|_2^2\right] \\ &= \mathbb{E}_{t,\mathbf{x}_0,\mathbf{y},\mathbf{z}}\left[\left\|\mathbf{s}_\theta\big(\big[\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t) + \sigma(t)\mathbf{z}\big], \mathbf{y}, t\big) + \frac{\mathbf{z}}{\sigma(t)}\right\|_2^2\right],\end{aligned} \tag{11}$$

using $\mathbf{x}_t = \boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},t) + \sigma(t)\mathbf{z}$, with $\mathbf{z} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. t is sampled uniformly in $[t_\epsilon, T]$ where t_ϵ is a minimal diffusion time used to avoid numerical instabilities.

3.3. Inference through reverse sampling

At inference time, we first sample an initial condition of the reverse process, corresponding to $\mathbf{x}_T$, with:

$$\mathbf{x}_T \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_T; \mathbf{y}, \sigma^2(T)\mathbf{I}), \tag{12}$$

This sample corresponds to corrupted speech $\mathbf{y}$ to which we add Gaussian noise with variance $\sigma(t)^2$, which approximates the training condition.

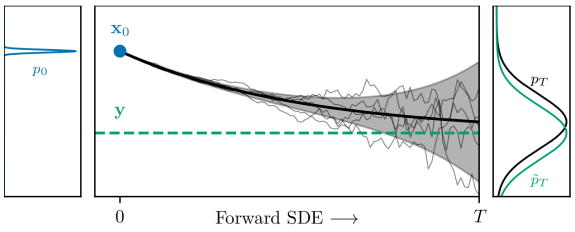


Fig. 1. *Visualization of the forward process* (6). *Mean curve is in solid black and variance is represented by the greyed area. Several realizations of the diffusion process are represented by thin black lines.*

Conditional generation is then performed by solving the following *plug-in reverse SDE* from $t = T$ to $t = 0$, where the score function is replaced by its estimator $\mathbf{s}_\theta$, assuming the latter was trained e.g. according to Section 3.2:

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, \mathbf{y}) + g(t)^2\, \mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t)\right] \mathrm{d}t + g(t)\, \mathrm{d}\bar{\mathbf{w}}, \qquad (13)$$

where $\mathbf{f}$ and g are the drift and diffusion terms defined in (6).

We use classical numerical solvers based on discretization of (13) according to a N points grid of the interval $[0,T]$. Since each reverse diffusion step calls the score network, the inference time of diffusion models is higher than their discriminative counterparts, by two orders of magnitude in our case. Fast inference schemes are discussed in the literature and are outside of the scope of this paper.

4. EXPERIMENTAL SETUP

4.1. Data

We use the WSJ0 corpus [29] for most experiments to ensure easier comparison between tasks. For comparison to bandwidth extension baselines, we use the VCTK corpus [30]. All data generation methods are accessible via our web page[2].

Speech Enhancement: The WSJ0+Chime dataset is generated using clean speech extracts from the Wall Street Journal corpus and noise signals from the CHiME3 dataset [31]. The mixture signal is created by randomly selecting a noise file and adding it to a clean utterance with a signal-to-noise ratio (SNR) sampled uniformly between 0 and 20dB.

Speech Dereverberation: The WSJ0+Reverb dataset is generated using clean speech data from the WSJ0 dataset and convolving each utterance with a simulated RIR. We use the PyRoomAcoustics engine [32] to simulate the RIRs. The reverberant room is modeled by sampling uniformly a target T_{60} between 0.4 and 1.0 seconds and room length, width and height in [5,15]×[5,15]×[2,6] m. A dry version of the room is created with the same geometric parameters with a fixed absorption coefficient of 0.99, to generate the corresponding anechoic target.

Bandwidth Extension: The WSJ0+BWR dataset is built with clean speech extracted from the WSJ0 corpus and a similar bandwidth reduction recipe as in [21, 23]. We pick an anti-aliasing filter type among Chebyshev, Butterworth, Elliptic and Bessel and a filter order among$\{2,4,8\}$. Decimating is then realized with a down-scaling factor sampled in$\{2,4,8\}$. The utterance is then resampled at the original 16 kHz with polyphase filtering. To compare against other baselines, we generate VCTK+BWR by replacing WSJ0 with VCTK as the base speech corpus, which we first resample to 16kHz, and use the same process as explained above.

4.2. Hyperparameters and training configuration

Data representation: Utterances are transformed using a short-time Fourier transform (STFT) with a window size of 510, a hop length of 128 and a Hann window. Square-root magnitude compression is carried on the spectrogram. For training, sequences of 256 STFT frames (i.e. 2s) are randomly extracted from the full-length utterances and normalized with respect to the corrupted mixture before being fed to the network.

Forward diffusion: Defined in (6), the stiffness parameter is fixed to γ=1.5, the extremal noise levels to $\sigma_{\min} = 0.05$ and $\sigma_{\max} = 0.5$. The minimal diffusion time defined in (11) is set to $t_\epsilon = 0.03$ as in [8].

Network architecture: The original architecture used for score estimation in [8] is the NCSN++ network proposed in [26]. NCSN++ is a multiresolution U-Net structure which includes in each layer a series of ResNet blocks using 2D convolutions, group normalization and fixed down/upsampling. Attention mechanism is used in the bottleneck, and the network leverages a parallel progressive growing path in addition to the skip connections. The noisy speech spectrogram $\mathbf{y}$ and the current diffusion process estimate $\mathbf{x}_t$ real and imaginary channels are stacked and fed to the network as input. The model is made noise-conditional by feeding each ResNet block with an encoded version of the current noise level $\sigma(t)$. More details about the architecture can be found in [8, 26]. For the generative model proposed in this paper, denoted as *SGMSE+M*, we use a lighter configuration of the NCSN++ architecture called NCSN++M . Ablation studies were designed to halve the number of parameters with almost no degradation, resulting in a network capacity of roughly 27.8M parameters. For the discriminative approach, denoted simply as *NCSN++M* in the following, the noise-conditioning layers are removed This ablation removes only 1.8% of the original number of parameters, which hardly modifies the network capacity.

Training configuration: We train the DNN for a maximum of 300 epochs using early stopping with a patience of 10 epochs. The generative approach SGMSE+M is trained with the denoising score matching criterion (11), and discriminative NCSN++M uses a simple mean-square error loss on the complex spectrogram. We use the Adam optimizer with a learning rate of 10^{-4} and an effective batch size of 16. We track an exponential moving average of the DNN weights with a decay of 0.999.

Inference: 50 time steps are used for reverse inference, adopting the predictor-corrector scheme [26] with one step of annealed Langevin dynamics correction.

4.3. Evaluation metrics

For instrumental evaluation of the speech enhancement and dereverberation performance, we use Perceptual Evaluation of Speech Quality (PESQ) [33] , extended short-term objective intelligibility (ESTOI) [34] and scale-invariant signal to distortion ratio (SI-SDR) [35] . For bandwidth extension we also include log spectral distance (LSD) as a common metric used in the literature. However, it must be stated that the aforementioned instrumental metrics may relate poorly with listening experiments, especially for bandwidth extension. We therefore complement our metrics benchmark with WV-MOS [23][3], which is a DNN-based mean opinion score (MOS) estimation, and was used by the authors for assessment of bandwidth extension performance. For comparability purposes to baselines on the VCTK corpus, we use regular STOI [36] instead of its extended version .

5. EXPERIMENTAL RESULTS AND DISCUSSION

5.1. Speech enhancement

In Table 1, we report speech enhancement performance on the WSJ0+Chime dataset. We notice that the generative SGMSE+M produces higher quality samples as measured by WV-MOS and PESQ.

[2]https://uhh.de/inf-sp-sgmsemultitask

[3]https://github.com/AndreevP/wvmos

Table 1. Results for denoising on WSJ0+Chime data.

Method	Type	WV-MOS	PESQ	ESTOI	SI-SDR
Mixture		1.44 ± 1.62	1.70 ± 0.49	0.78 ± 0.14	10.0 ± 5.7
NCSN++M	D	3.65 ± 0.48	2.67 ± 0.69	$\mathbf{0.93 \pm 0.06}$	$\mathbf{19.5 \pm 4.4}$
SGMSE+M	G	$\mathbf{3.77 \pm 0.32}$	$\mathbf{2.94 \pm 0.60}$	0.92 ± 0.06	18.0 ± 5.1

Table 2. Results for dereverberation on WSJ0+Reverb data.

Method	Type	WV-MOS	PESQ	ESTOI	SI-SDR
Mixture		1.78 ± 0.99	1.36 ± 0.19	0.46 ± 0.12	-7.3 ± 5.5
NCSN++M	D	2.96 ± 0.38	2.19 ± 0.48	$\mathbf{0.87 \pm 0.05}$	$\mathbf{7.2 \pm 3.7}$
SGMSE+M	G	$\mathbf{3.43 \pm 0.33}$	$\mathbf{2.64 \pm 0.42}$	$\mathbf{0.87 \pm 0.05}$	6.4 ± 4.2

Table 3. Results for bandwidth extension on WSJ0+BWR data.

Method	Type	WV-MOS ↑	ESTOI ↑	LSD ↓
Mixture		2.45 ± 1.01	0.72 ± 0.21	2.31 ± 0.32
AE-NCSN++M	D	2.17 ± 0.93	0.71 ± 0.19	1.81 ± 0.21
NCSN++M	D	2.25 ± 0.87	0.73 ± 0.16	2.21 ± 0.30
SGMSE+M	G	$\mathbf{3.43 \pm 0.48}$	$\mathbf{0.83 \pm 0.13}$	$\mathbf{1.44 \pm 0.17}$

It is however slightly outperformed by discriminative NCSN++M on intelligibility and noise removal. Indeed, in a denoising task, the interference does not share any information with the target speech, making it relatively easy for a discriminative approach to remove the interference without distorting the target. However, we show in the uploaded listening examples that the discriminative approach tends to destroy low-energy speech regions for low SNRs, whereas the generative model does not. A larger benefit of the generative approach is observed when training and testing data have a stronger mismatch [8].

5.2. Speech dereverberation

In Table 2, we report dereverberation results on the WSJ0+Reverb dataset. Here, generative SGMSE+M clearly outperforms discriminative NCSN++M in terms of quality by a large margin on WV-MOS and PESQ, and performs on par on ESTOI and SI-SDR. For dereverberation, in contrast to denoising, the interference model is completely dependent on the target as it is a filtered version of the latter (Eq. (2)). The generative model is able to extract the speech cues and directly reconstructs it with very little reverberation. The discriminative method, however, cannot do so without introducing significant distortions.

5.3. Bandwidth extension

Results on WSJ0+BWR: In Table 3 we report bandwidth extension performance on the WSJ0+BWR dataset. Interestingly, using a STFT representation did not allow the discriminative approach to recreate the lost high-frequency content. The approach simply learnt an identity mapping, and similar results were observed when experimenting with other STFT-based DNN backbones and data. For this discriminative case, we modified the NCSN++M architecture to use a learnt encoder and decoder, as in e.g. [19] . The resulting approach, denoted as AE-NCSN++M in the following, uses a single 1D convolutional layer with 256 filters of length 510 and stride 128, so that the learnt representation is equivalent to the chosen STFT filterbank. As opposed to NCSN++M, AE-NCSN++M is able to generate high-frequency components, however the reconstruction quality is overall poor, which is to be expected given the generative nature of the bandwidth extension task. By learning an approximate identity mapping, NCSN++M performs better than AE-NCSN++M on instrumental metrics although it does not actually perform bandwidth extension.

Table 4. Results for bandwidth extension on VCTK data. * means that the results were taken from [23]. † means that the method was trained on each bandwidth reduction factor separately.

Bandwidth	Type	1kHz		2kHz		4kHz	
		WV-MOS	STOI	WV-MOS	STOI	WV-MOS	STOI
Mixture		1.36	0.79	2.34	0.89	3.52	0.99
TUNet† [20]	D	-	-	-	-	3.86	0.98
TFiLM*† [19]	D	1.65	0.81	2.27	0.91	3.49	**1.00**
HiFi++*† [23]	G	**3.71**	**0.86**	**3.95**	**0.94**	4.16	**1.00**
VoiceFixer [21]	G	2.50	0.73	3.35	0.78	3.81	0.83
NuWave2 [24]	G	-	-	-	-	3.76	0.97
SGMSE+M	G	3.25	0.83	3.70	0.93	**4.20**	**1.00**

In contrast, generative SGMSE+M performs much better in all metrics, generating plausible content even when the Nyquist frequency is down to 2kHz. When the Nyquist frequency is down to 1kHz, the approach can struggle with generating the right consonants in some cases. Typically the generation process may mistake $[ch]$ for $[s]$ as the information needed to differentiate those sounds is available only at frequencies way above 1kHz. Integrating a linguistic or visual model could be here envisaged to make the approach robust to this lack of acoustic cues.

Results on VCTK+BWR: In Table 4, we compare the proposed generative SGMSE+M on the VCTK+BWR test set against HiFI++ [23], VoiceFixer[4] [21], TFiLM [19], TUNet[5] [20] and NuWave2[6] [24]. Please note that HiFi++, TFiLM and TUNet are trained on each input bandwidth separately, while our generative model SGMSE+M as well as VoiceFixer and NuWave2 are bandwidth-agnostic. We use the official implementations for all approaches without retraining, except HiFi++ as no code is available, and TFiLM as no multi-speaker model is provided. When a method is not trained to restore speech at 16kHz, we use it at the nominal sampling rate then downsample its output to 16kHz. SGMSE+M achieves on par results with HiFi++ on 4kHz bandwidth and worse on 1kHz and 2kHz bandwidths, which is partially due to the fact that HiFi++ is trained separately for each input bandwidth. Using neural vocoders incorporating speech knowledge, as is the case for HiFi++, also probably helps improve robustness to very low input bandwidths. Against all other approaches than HiFi++, SGMSE+M performs significantly better on almost all metrics and conditions.

6. CONCLUSION

The goal of this work is to analyse the potential benefit of recent diffusion-based generative approaches against discriminative approaches on various speech restoration tasks. For this, we apply our recently proposed diffusion generative model to speech enhancement, dereverberation and bandwidth extension, and compare against a discriminative approach using the same DNN architecture . We observe that the generative approach performs globally better than its discriminative counterpart on all tasks, with the strongest benefit for non-additive distortion models, like in dereverberation and bandwidth extension. Furthermore, we show that the proposed bandwidth-agnostic method performs slightly worse or on par in comparison with a recent bandwidth-dependent approach, and largely outperforms other discriminative and bandwidth-agnostic generative approaches.

[4]https://github.com/haoheliu/voicefixer

[5]https://github.com/NXTProduct/TUNet

[6]https://github.com/mindslab-ai/nuwave2

7. REFERENCES

[1] T. Gerkmann and E. Vincent, *Spectral Masking and Filtering*. John Wiley & Sons, 2018.

[2] 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.

[3] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[4] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Int. Conf. Machine Learning (ICML)*, Apr. 2015.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Neural Information Proc. Systems (NIPS)*, Dec. 2020.

[6] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Neural Information Proc. Systems (NIPS)*, Dec. 2019.

[7] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Interspeech*, Sept. 2022.

[8] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *arXiv*, 2022.

[9] 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)*, June 2022.

[10] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv*, 2022.

[11] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," June 2021.

[12] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *Asia-Pacific Signal and Information Processing Association (APSIPA)*, Dec. 2021.

[13] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, vol. 59. Springer, 2011.

[14] E. Habets, *Single- and Multi-Microphone Speech Dereverberation Using Spectral Enhancement*. PhD thesis, 2007.

[15] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2014.

[16] T. Gerkmann, "Cepstral weighting for speech dereverberation without musical noise," in *Proc. Euro. Signal Proc. Conf. (EUSIPCO)*, Sept. 2011.

[17] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 25, no. 7, pp. 1492–1501, 2017.

[18] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *Proc. Euro. Signal Proc. Conf. (EUSIPCO)*, Sept. 2019.

[19] S. Birnbaum, V. Kuleshov, Z. Enam, P. W. Koh, and S. Ermon, "Temporal film: Capturing long-range sequence dependencies with feature-wise modulations," in *Neural Information Proc. Systems (NIPS)*, Dec. 2019.

[20] V.-A. Nguyen, A. H. T. Nguyen, and A. W. H. Khong, "TUNet: A block-online bandwidth extension model based on transformers and self-supervised pretraining," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, June 2022.

[21] H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "Voicefixer: Toward general speech restoration with neural vocoder," *arXiv*, 2021.

[22] H. Liu, W. Y. Choi, X. Liu, Q. Kong, Q. Tian, and D. Wang, "Neural vocoder is all you need for speech super-resolution," in *Interspeech*, Sept. 2022.

[23] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: a unified framework for bandwidth extension and speech enhancement," *arXiv*, 2022.

[24] S. Han and J. Lee, "Nu-wave 2: A general neural audio upsampling model for various sampling rates," in *Interspeech*, Sept. 2022.

[25] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, vol. 82. Journal of the American Statistical Association, 2000.

[26] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Int. Conf. Learning Repr. (ICLR)*, May 2021.

[27] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[28] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching.," *Journal of Machine Learning Research*, vol. 6, p. 695–709, 2005.

[29] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete," May 2007.

[30] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Speech Synthesis Workshop (SSW)*, Sept. 2016.

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

[32] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Apr. 2018.

[33] 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.

[34] J. Jensen and C. 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.

[35] J. L. 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.

[36] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 19, no. 7, pp. 2125–2136, 2011.

4.3 Combining Predictive Approaches and Diffusion-based Generative Models for Speech Enhancement and Dereverberation [P8]

Abstract

Diffusion models have shown a great ability at bridging the performance gap between predictive and generative approaches for speech enhancement. We have shown that they may even outperform their predictive counterparts for non-additive corruption types or when they are evaluated on mismatched conditions. However, diffusion models suffer from a high computational burden, mainly as they require to run a neural network for each reverse diffusion step, whereas predictive approaches only require one pass. As diffusion models are generative approaches they may also produce vocalizing and breathing artifacts in adverse conditions. In comparison, in such difficult scenarios, predictive models typically do not produce such artifacts but tend to distort the target speech instead, thereby degrading the speech quality. In this work, we present a stochastic regeneration approach where an estimate given by a predictive model is provided as a guide for further diffusion. We show that the proposed approach uses the predictive model to remove the vocalizing and breathing artifacts while producing very high quality samples thanks to the diffusion model, even in adverse conditions. We further show that this approach enables to use lighter sampling schemes with fewer diffusion steps without sacrificing quality, thus lifting the computational burden by an order of magnitude. Source code and audio examples are available online (this https URL).

Reference

Jean-Marie Lemercier, Julius Richter, Simon Welker and Timo Gerkmann "StoRM: A Diffusion-based Stochastic Regeneration Model for Speech Enhancement and Dereverberation", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 31, pp. 2724-2737, 2023, DOI: 10.1109/TASLP.2023.3294692

Copyright Notice

Authors' Contributions

Jean-Marie Lemercier is the first author of this publication. He implemented all algorithms, trained the neural networks used in the paper, conducted the experimental validation, and wrote the manuscript. Julius Richter and Simon Welker brought their feedback on all methods through discussions and also helped with reviewing the manuscript. Timo Gerkmann brought insights on the experimental validation, mathematical derivations, and reviewed the manuscript.

StoRM: A Diffusion-based Stochastic Regeneration Model for Speech Enhancement and Dereverberation

Jean-Marie Lemercier, *Student Member, IEEE*, Julius Richter, *Student Member, IEEE*, Simon Welker, *Student Member, IEEE*, Timo Gerkmann, *Senior Member, IEEE*

***Abstract*—Diffusion models have shown a great ability at bridging the performance gap between predictive and generative approaches for speech enhancement. We have shown that they may even outperform their predictive counterparts for non-additive corruption types or when they are evaluated on mismatched conditions. However, diffusion models suffer from a high computational burden, mainly as they require to run a neural network for each reverse diffusion step, whereas predictive approaches only require one pass. As diffusion models are generative approaches they may also produce vocalizing and breathing artifacts in adverse conditions. In comparison, in such difficult scenarios, predictive models typically do not produce such artifacts but tend to distort the target speech instead, thereby degrading the speech quality. In this work, we present a stochastic regeneration approach where an estimate given by a predictive model is provided as a guide for further diffusion. We show that the proposed approach uses the predictive model to remove the vocalizing and breathing artifacts while producing very high quality samples thanks to the diffusion model, even in adverse conditions. We further show that this approach enables to use lighter sampling schemes with fewer diffusion steps without sacrificing quality, thus lifting the computational burden by an order of magnitude. Source code and audio examples are available online[1].**

***Index Terms*—score-based generative models, diffusion models, speech enhancement, speech dereverberation, predictive learning.**

I. Introduction

In real-life scenarios and modern communication devices, clean speech sources are often polluted by background noise, interfering speakers, room acoustics and codec degradation [1], [2]. We refer to this phenomenon as *speech corruption*, and denote by *speech restoration* the art of recovering clean speech from the corrupted signal [3]. On the one hand, traditional speech restoration methods leverage the statistical properties of the target and interference signals in various domains e.g. time, spectrum, cepstrum or spatial distribution [4]. On the other hand, machine learning techniques try to learn these statistical properties and how to exploit them from data [5]. Machine learning algortihms can be categorized into predictive (also called discriminative) approaches and generative approaches. We will choose the term *predictive* over *discriminative* as it fits both classification and regression tasks [6]. The field of speech restoration is dominated by predictive approaches that use supervised learning to learn a single best deterministic mapping between corrupted speech $\mathbf{y}$ and the corresponding clean speech target $\mathbf{x}$ [5]. These methods include for instance time-frequency (T-F) masking [7], time domain methods [8], [9] or direct spectro-temporal mapping [10]. They have contributed to drastically increasing the quality of speech restoration algorithms. However, they can distort target speech and suffer from generalizability issues [11], [12].

In contrast, generative models implicitly or explicitly learn the target distribution and allow to generate multiple valid estimates instead of a single best estimate as for predictive approaches [6]. Generative approaches include variational auto-encoders (VAEs) learning explicit density estimations [13]–[16], normalizing flows adding invertible transforms to obtain tractable marginal likelihoods [17], [18], generative adversarial networks (GANs) estimating implicit distributions [19], [20] and diffusion approaches [21]–[23]. We talk of *conditional* generative models when a covariate $\mathbf{c}$ is used to guide the generation, leading to the conditional distribution $p(\mathbf{x}|\mathbf{c})$ [6]. This conditioning can either be another modality describing the data (e.g. $\mathbf{c}$ could be video when $\mathbf{x}$ is speech), or a modified version of the data, an obvious example being corrupted speech $\mathbf{y}$ when the underlying task is speech restoration. By integrating stochasticity in their latent structure, generative models can capture the inherent uncertainty of the data distribution and produce realistic samples belonging to that distribution rather than a mean of optimal candidates [6]. In doing so, they may obtain better perceptual metrics at the cost of higher point-wise distortion [24]. In the imaging domain, it was observed that predictive approaches tend to brush over the fine-grained details of the considered domain [24], [25]. Furthermore, predictive models may result in limited generalization abilities towards unseen noise types or speakers as compared to generative models, which is demonstrated for diffusion-based generative speech enhancement in [12].

We focus in this work on such diffusion-based generative models, or simply *diffusion models*, which have met great

This work has been funded by the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380, the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169) and DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the Grant-No. HIDSS-0002.

Simon Welker is with the Signal Processing Group, Department of Informatics, Universität Hamburg, 22527 Hamburg Germany, and with the Center for Free-Electron Laser Science, DESY, 22607 Hamburg, Germany (e-mail: simon.welker@uni-hamburg.de).

The other authors are with the Signal Processing Group, Department of Informatics, Universität Hamburg, 22527 Hamburg Germany (e-mail: {jeanmarie.lemercier; julius.richter; timo.gerkmann}@uni-hamburg.de).

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

success in generating high-quality samples of natural images [21]–[23], [26]. Diffusion models use a *forward process* to slowly turn data into a tractable prior, usually a standard normal distribution, and train a neural network to solve the *reverse process* to generate clean data from this prior [27]. These diffusion models can also be used for conditional generation in restoration tasks, which has recently been proposed for speech processing tasks such as enhancement and dereverberation [12], [28]–[30] as well as bandwidth extension [11], [31].

One limiting aspect of diffusion models is their heavy computational burden. Several steps are needed for reverse diffusion, each of them calling the neural network used for score estimation. Much effort has been recently put into reducing this number of steps, either by optimization of the reverse noise schedule [32], modifications in the formulation of the diffusion processes [33], [34], or projection into a latent space [35] or a reduced subspace [36]. We also observed in past experiments that our previously proposed diffusion model is prone to confuse phonemes and generate vocalizing artifacts when facing very adverse conditions. This is due to the generative behaviour of the model under high uncertainty over the presence or nature of speech, and this naturally leads to a degradation, e.g. in automatic speech recognition (ASR).

In this work, we propose a *stochastic regeneration* scheme combining predictive and generative models to produce high quality samples while reducing the computational burden of diffusion models and their tendency to generate unwanted artifacts. We propose to first use a predictive approach to estimate a restored version of the corrupted speech. This estimate is then used as a guide by a diffusion model, which requires only a few diffusion steps to output a final clean speech estimation where the distortions introduced by the predictive stage are corrected while vocalizing artifacts and phonetic confusions are avoided. Both listening experiments and instrumental metrics confirm an impressive state-of-the-art perceptual quality of our proposed approach. Other refinement approaches using diffusion models were recently proposed. The *stochastic refinement* approach [24], [37] subtracts the output of the predictive model from the corrupted speech, and this residual is used for further estimation by a diffusion model. We argue hereafter that learning the residual is however a hard task and demonstrate that our approach outperforms this stochastic refinement in terms of instrumentally measured speech quality. Another refinement approach using diffusion models is *denoising diffusion restoration models* [38]–[40], where the corruption operator is assumed to be known (or at least its singular value decomposition) and is used to modify the reverse diffusion process at inference time.

We evaluate our proposed approach for speech enhancement with low input signal-to-noise ratios (SNRs) and speech dereverberation, using clean speech from the WSJ0 corpus [41]. We also show ASR results on the TIMIT dataset [42], and report results on the standardized Voicebank/DEMAND dataset [43]. Ablation studies are performed on sampling efficiency, intial predictor mismatch and training strategy.

II. Score-based Diffusion Models

Diffusion models originally use discrete-time diffusion processes modeled by Markov chains [22]. They have been recently extended to continuous-time diffusion processes formulated by stochastic differential equations (SDEs) in [44], allowing for new training paradigms such as score matching [45], [46]. This class model is subsequently denoted as *score-based diffusion models*. Score-based diffusion models are defined by three components: a forward diffusion process, a score function estimator, and a sampling method for inference.

A. Forward and reverse processes

The stochastic forward process $\{\mathbf{x}_\tau\}_{\tau=0}^{T}$ used in score-based diffusion models is defined as an Itô SDE [44], [47]:

$$\mathrm{d}\mathbf{x}_\tau = \mathbf{f}(\mathbf{x}_\tau, \tau)\mathrm{d}\tau + g(\tau)\mathrm{d}\mathbf{w}, \tag{1}$$

where $\mathbf{x}_\tau$ is the current state of the process indexed by $\tau \in [0, T]$ with the initial condition representing clean speech $\mathbf{x}_0 = \mathbf{x}$. The continuous *diffusion time* variable τ relates to the progress of the stochastic process and should not be mistaken for our usual notion of *signal time*. As our process is defined in the complex spectrogram domain, independently for each T-F bin, the variables in bold are vectors in $\mathbb{C}^d$ containing the coefficients of a flattened complex spectrogram— with d the product of the time and frequency dimensions— whereas variables in regular font represent real scalar values. The set $\{\mathbf{x}_\tau\}_{\tau\in]0,T[}$ can be seen as latent variables used to parameterize the conditional distribution $p(\mathbf{x}_\tau|\mathbf{x}_0, \mathbf{y})$. The stochastic process $\mathbf{w}$ denotes a standard d-dimensional Brownian motion, that is, $\mathrm{d}\mathbf{w}$ is a zero-mean Gaussian variable with standard deviation $\mathrm{d}\tau$ for each T-F bin.

The *drift* function $\mathbf{f}$ and *diffusion* coefficient g as well as the initial condition $\mathbf{x}_0$ and the final diffusion time T uniquely define the Itô process $\{\mathbf{x}_\tau\}_{\tau=0}^{T}$ [47]. Under some regularity conditions on $\mathbf{f}, g$ allowing a unique and smooth solution to the Kolmogorov equations associated to (1), the reverse process $\{\mathbf{x}_\tau\}_{\tau=T}^{0}$ is another diffusion process defined as the solution of the following SDE [44], [48]:

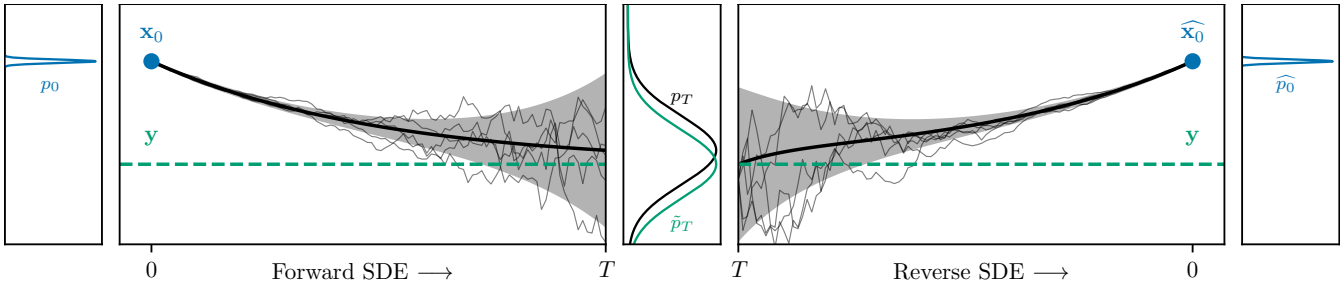
$$\mathrm{d}\mathbf{x}_\tau = \left[-\mathbf{f}(\mathbf{x}_\tau, \tau) + g(\tau)^2\nabla_{\mathbf{x}_\tau}\log p_\tau(\mathbf{x}_\tau)\right]\mathrm{d}\tau + g(\tau)\mathrm{d}\bar{\mathbf{w}}, \tag{2}$$

where $\mathrm{d}\bar{\mathbf{w}}$ is a d-dimensional Brownian motion for the time flowing in reverse and $\nabla_{\mathbf{x}_\tau}\log p_\tau(\mathbf{x}_\tau)$ is the *score function*, i.e. the gradient of the logarithm data distribution for the current process state $\mathbf{x}_\tau$.

Speech restoration tasks can be regarded either as one-to-one mapping tasks between corrupted speech $\mathbf{y}$ and $\mathbf{x}_0$, which leads to predictive modelling; or as conditional generation tasks, i.e. generation of $\mathbf{x}_0$ conditioned on $\mathbf{y}$. Previous diffusion-based approaches proposed to condition the process explicitly within the neural network [49] or through guided classification [26]. In [28], the conditioning is directly incorporated into the diffusion process by defining the forward process as the solution to the following SDE:

$$\mathrm{d}\mathbf{x}_\tau = \underbrace{\gamma(\mathbf{y} - \mathbf{x}_\tau)}_{:=\mathbf{f}(\mathbf{x}_\tau,\mathbf{y})}\mathrm{d}\tau + \underbrace{\left[\sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\tau}\sqrt{2\log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}\right]}_{:=g(\tau)}\mathrm{d}\mathbf{w}. \tag{3}$$

Fig. 1: *Visualization of the forward and backward processes in* (3)*. Mean curve* (5) *is in solid black and variance* (6) *is represented by the greyed area. Several realizations of the diffusion process are represented by thin black lines. The mismatch between p_τ centered on $\mathbf{x}_\tau$ and $\tilde{p}_\tau$ centered on $\mathbf{y}$ comes from the fact that the mean in* (5) *can not reach $\mathbf{y}$ in finite time. This mismatch causes unavoidable bias in the reverse process, even were the score perfectly known.*



This equation belongs to the class of Ornstein-Uhlenbeck SDEs [47], a subclass of Itô SDEs in which the drift function $\mathbf{f}$ is affine in $\mathbf{x}_\tau$ and does not depend on τ, and the diffusion coefficient g only depends on τ. The equation introduces a *stiffness* hyperparameter γ controlling the slope of the decay from $\mathbf{y}$ to $\mathbf{x}_0$, and σ_{min} and σ_{max} are two hyperparameters controlling the *noise scheduling*, that is, the amount of Gaussian white noise injected at each timestep of the process.

The interpretation of our forward process in Eq. (3), visualized on Fig. 1, is as follows: at each time step and for each T-F bin independently, an infinitesimal amount of corruption is added to the current process state $\mathbf{x}_\tau$, along with Gaussian noise with standard deviation $g(\tau)\mathrm{d}\tau$. Therefore, the mean of the current process decays exponentially towards $\mathbf{y}$ while the variance increases as in the variance-exploding scheme of Song et al. [44], leading to a final distribution $\mathbf{x}_\tau$ which is the corrupted signal $\mathbf{y}$ with some additional Gaussian noise. Given an initial condition $\mathbf{x}_0$ and the covariate $\mathbf{y}$, the solution to (3) admits the following complex Gaussian distribution for the process state $\mathbf{x}_\tau$ called *perturbation kernel*:

$$p_{0,\tau}(\mathbf{x}_\tau|\mathbf{x}_0,\mathbf{y}) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_\tau; \boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},\tau), \sigma(\tau)^2\mathbf{I}\right), \quad (4)$$

Following [50], we determine closed-form solutions for the mean $\boldsymbol{\mu}$ and variance $\sigma(\tau)^2$:

$$\boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},\tau) = \mathrm{e}^{-\gamma\tau}\mathbf{x}_0 + (1-\mathrm{e}^{-\gamma\tau})\mathbf{y}\,, \quad (5)$$

$$\sigma(\tau)^2 = \frac{\sigma_{\text{min}}^2\left(\left(\sigma_{\text{max}}/\sigma_{\text{min}}\right)^{2\tau} - \mathrm{e}^{-2\gamma\tau}\right)\log(\sigma_{\text{max}}/\sigma_{\text{min}})}{\gamma + \log(\sigma_{\text{max}}/\sigma_{\text{min}})}\,. \quad (6)$$

B. Score function estimator

When performing inference, one tries to solve the reverse SDE in Eq. (2). In the general case, the score function $\nabla_{\mathbf{x}_\tau}\log p_\tau(\mathbf{x}_\tau)$ is not readily available, it can however be estimated by a deep neural network (DNN) $\mathbf{s}_\phi$ called the *score model*. Given the simple Gaussian form of the perturbation kernel $p_{0,\tau}(\mathbf{x}_\tau|\mathbf{x}_0,\mathbf{y})$ (4) and the regularity conditions exhibited by the mean and variance, a *denoising score matching* objective can be used to train the score model $\mathbf{s}_\phi$ [45], [46].

The score function of the perturbation kernel is:

$$\nabla_{\mathbf{x}_\tau}\log p_{0,\tau}(\mathbf{x}_\tau|\mathbf{x}_0,\mathbf{y}) = -\frac{\mathbf{x}_\tau - \boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},\tau)}{\sigma(\tau)^2}. \quad (7)$$

Once a clean utterance $\mathbf{x}_0$ and noisy utterance $\mathbf{y}$ are picked in the training set, the current process state is obtained as $\mathbf{x}_\tau = \boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},\tau) + \sigma(\tau)\mathbf{z}$, with $\mathbf{z}\sim\mathcal{N}_{\mathbb{C}}\left(\mathbf{z};\mathbf{0},\mathbf{I}\right)$. We can therefore write the denoising score matching objective as follows [44]:

$$\mathcal{J}^{(\text{DSM})}(\phi) = \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{y}),\mathbf{z},\mathbf{x}_\tau}\left[\left\|\mathbf{s}_\phi(\mathbf{x}_\tau,\mathbf{y},\tau) + \frac{\mathbf{z}}{\sigma(\tau)}\right\|_2^2\right]. \quad (8)$$

Here, we sample τ sampled uniformly in $[\tau_\epsilon, T]$ where τ_ϵ is a minimal diffusion time used to avoid numerical instabilities. This approach is analogous to the denoising objective used in the discrete-time formulation by [22], where one estimates the noise added at each step to learn the reverse process.

C. Inference through reverse sampling

At inference time, we first sample an initial condition of the reverse process, corresponding to $\mathbf{x}_T$, with:

$$\mathbf{x}_T \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_T; \mathbf{y}, \sigma^2(T)\mathbf{I}), \quad (9)$$

This sample only approximates the training condition, as the final process distribution $p_T(\mathbf{x}_T)$ does not perfectly match $p(\mathbf{y})$ (see Fig. 1).

Conditional generation is then performed by solving the so-called *plug-in reverse SDE* from $\tau = T$ to $\tau = 0$, where the score function is replaced by its estimator $\mathbf{s}_\phi$, assuming the latter was trained e.g. according to Section II-B:

$$\mathrm{d}\mathbf{x}_\tau = \left[-\mathbf{f}(\mathbf{x}_\tau,\mathbf{y}) + g(\tau)^2\mathbf{s}_\phi(\mathbf{x}_\tau,\mathbf{y},\tau)\right]\mathrm{d}\tau + g(\tau)\mathrm{d}\bar{\mathbf{w}} \quad (10)$$

We use classical numerical solvers based on a discretization of (10) according to a uniform grid of N points on the interval $[0,T]$ (no minimal diffusion time is needed here). Classical solvers include the Euler-Maruyama method, higher-order single-step methods, and predictor-corrector sampler schemes [44]. In the latter, at each reverse step τ, the predictor uses a single-step method like Euler-Maruyama to generate $\mathbf{x}_\tau$, and the corrector uses the output of the score network to ensure consistency of the resulting sample with a marginal distribution consistent with the score estimate.

For notational convenience, we will denote by G_ϕ the generative model corresponding to the reverse diffusion process solver parameterized by the plug-in SDE (10) and the score network $\mathbf{s}_\phi$, such that the final estimate is $\widehat{\mathbf{x}} = G_\phi(\mathbf{y})$.

III. Stochastic Regeneration with Diffusion Models

A. Predictive artifacts for images and spectrograms

Most problems in speech restoration (e.g. denoising without knowledge of the environmental noise signal, dereverberation or bandwidth extension) are ill-posed inverse problems. This means that either (i) it is not possible to exactly retrieve $\mathbf{x}$ given $\mathbf{y}$ (e.g. dereverberation with a known non-minimum-phase room impulse response); or (ii) many versions of clean speech $\mathbf{x}^{(i)}$ can correspond to the same corrupted speech $\mathbf{y}$ (e.g. bandwidth extension). Consider a predictive model D_θ trained with a L^2 regression objective $\mathbb{E}_{(\mathbf{x},\mathbf{y})}||\mathbf{x} - D_\theta(\mathbf{y})||_2^2$. Because of the expectation over all training examples, an optimal predictive model learns the mapping to the posterior mean $\mathbf{y} \to \mathbb{E}[\mathbf{x}|\mathbf{y}]$, thereby minimizing the *average* distortion over all training examples. This phenomenon is known as *regression to the mean* [6], [24]. This can be problematic if the posterior distribution $p(\mathbf{x}|\mathbf{y})$ has an intricate structure and is thus not well represented by its mean $\mathbb{E}[\mathbf{x}|\mathbf{y}]$ (see Figure 5 here, or Figure 1 in [51]).

In image processing problems, this translates to predictive approaches being incapable of reproducing fine-grained details like e.g. edges and hair structure in human portraits [24], [25]. Our interpretation is that, given that these regions have the highest variability across natural image data [52], mapping directly to the posterior mean $\mathbb{E}[\mathbf{x}|\mathbf{y}]$ will smooth out these details by mistaking them for noise. Therefore the predictive model will output a sample which does not necessarily lie on the posterior data manifold. When training predictive models for spectrogram estimation, we also observe that the models tend to introduce distortions in the target speech when the corruption level is high, leading to *overdenoising* effects and loss of output resolution [11], [53].

The link between these observations in the imaging and speech domains is the following: performing speech restoration in the spectrogram domain can be seen as image processing (with two pixel dimensions representing the real and imaginary parts instead of three for RGB image processing). The distortions observed in the predictive model output are a smoothing effect *in the spectrogram seen as an image*. This results in removal of fine-grained detail corresponding to quiet regions (i.e. low luminosity detail in images) and onsets or offsets (i.e. edges in images) in the spectrogram. This is visualized in the third column of Fig. 2, where we directly compare spectrograms from our previous study [11] with images in Welker et al. [25, Fig. 1] reproduced here.

Several paradigms using generative modelling can be envisaged to correct this bias of the predictive model without having to resort to a full-fletched computationally heavy diffusion-based generative model. Next, we present two of these approaches, namely *stochastic refinement* by Whang et al. [24] and *stochastic regeneration* which we propose here.

B. Stochastic refinement

Instead of solving the reverse diffusion process from noisy speech to a clean speech estimate, the *stochastic refinement* approach by Whang et al. [24] uses both a predictive approach and a generative diffusion model for efficient inference.

A predictive model D_θ serves as an *initial predictor* producing an estimate $D_\theta(\mathbf{y})$. This estimate often lacks fine-grained detail and has significant target speech distortions, especially for corruption models like reverberation [11]. Let us write the predictive model output as:

$$D_\theta(\mathbf{y}) = \mathbf{x} - \mathbf{x}^{(\mathrm{dis})} + \tilde{\mathbf{n}}. \tag{11}$$

The *target distortion* $\mathbf{x}^{(\mathrm{dis})}$ is the artifact introduced by the predictive model: it contains target cues that were mistaken for corruption by the model, and consequently distorted. The residual corruption $\tilde{\mathbf{n}}$ is what remains of the interference (e.g. noise or reverberation) after being processed by the model. There is behind this decomposition an underlying orthogonality assumption between $\mathbf{x}$ and $\mathbf{n}$, which implies orthogonality between $\mathbf{x}^{(\mathrm{dis})}$ and $\tilde{\mathbf{n}}$ [54].

A diffusion-based generative model G_ϕ is then used to learn the distribution of the *ideal residue* $\mathbf{r}_\mathbf{x} = \mathbf{x} - D_\theta(\mathbf{y})$, starting from the *noisy residue* $\mathbf{r}_\mathbf{y} = \mathbf{y} - D_\theta(\mathbf{y})$. Finally, the ideal residue estimate is added to the predictor estimate:

$$\widehat{\mathbf{x}} = D_\theta(\mathbf{y}) + \widehat{\mathbf{r}_\mathbf{x}} \tag{12}$$

$$= D_\theta(\mathbf{y}) + G_\phi(\mathbf{y} - D_\theta(\mathbf{y})) \tag{13}$$

Results in [24], [37] seem to indicate that this stochastic refinement approach performs as expected, outperforming the initial predictor on perceptual metrics and the pure generative approach with fewer diffusion steps. However, we argue that learning the residual is suboptimal as the residual data distribution $p(\mathbf{r}_\mathbf{x})$ does not have a structure like the target data distribution $p(\mathbf{x})$. Indeed, using (11), one can rewrite $\mathbf{r}_\mathbf{x}$ as:

$$\begin{aligned}\mathbf{r}_\mathbf{x} &= \mathbf{x} - D_\theta(\mathbf{y}) \\ &= \mathbf{x}^{(\mathrm{dis})} - \tilde{\mathbf{n}},\end{aligned} \tag{14}$$

and notice that the distribution of $\mathbf{r}_\mathbf{x}$ highly depends on the choice of the predictive model as well as on the task, which does not assure a structured distribution in the general case. We show examples in Fig. 3 of residuals generated by predictive models for speech enhancement and dereverberation, which confirm this observation. For dereverberation (similarly for deblurring, shown in [24]), the residue has an overall structure somewhat similar to the target, because of the convolutional corruption model. However, the formants structure is severely degraded. For denoising, the residue has no clear structure (compared to e.g. clean speech) and we thus argue that it cannot be easily estimated by a generative process. In Whang et al. [24], it is shown that the residual distribution has lower entropy per pixel than the original distribution, which makes learning the residual easier. This is also true for our denoising and dereverberation tasks. However the pointwise entropy of a distribution relates to the quantity of information that needs to be learnt by the model and does not capture global structures in the data which can actually help facilitate training. Most

Fig. 2: *Visualization of samples obtained with predictive approach (NCSN++M, see Section IV) and generative model (SGMSE+M, see [12] and Section IV) for two ill-posed problems, namely speech dereverberation (top, from [11]) and JPEG artifact removal (bottom, from [25]). Spectrograms horizontal and vertical axes represent time and frequency respectively.*

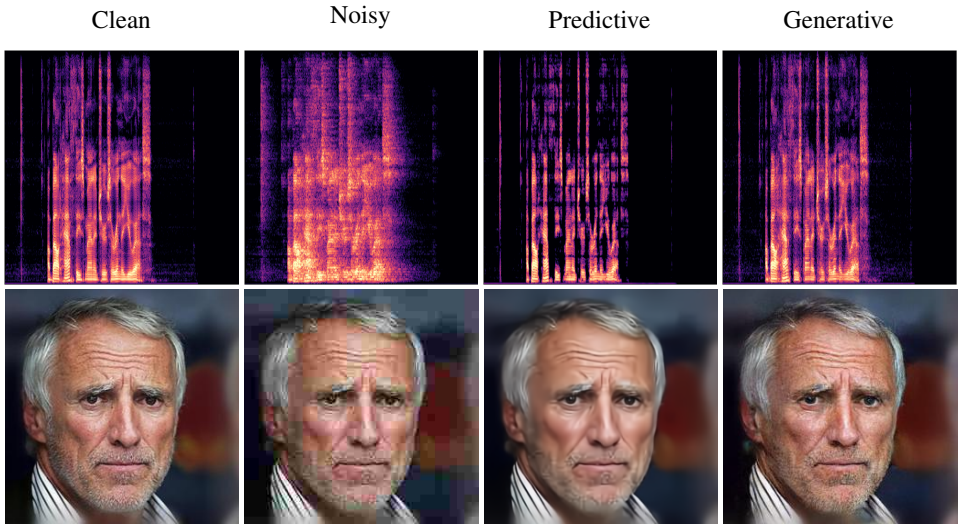
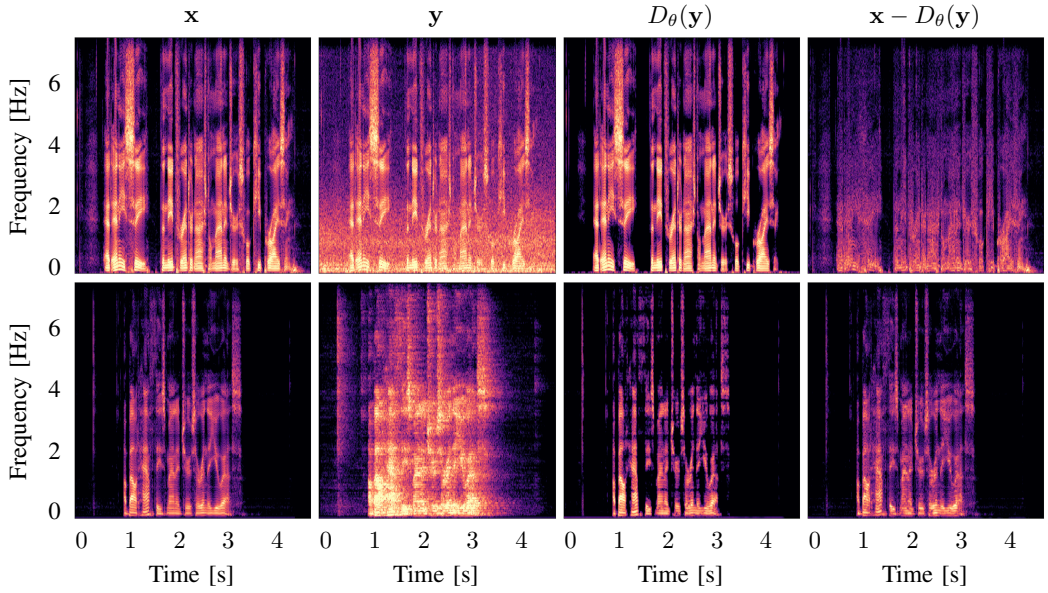


Fig. 3: *Log-energy spectrograms of clean, noisy, processed and residual utterances for denoising (top) and dereverberation (bottom). The predictor used is NCSN++M .*



importantly, when rewriting the noisy residue $\mathbf{r_y}$ as:

$$\begin{aligned}\mathbf{r_y} &= \mathbf{y} - D_\theta(\mathbf{y}) \\ &= \mathbf{x}^{(\text{dis})} + \mathbf{n} - \tilde{\mathbf{n}},\end{aligned} \tag{15}$$

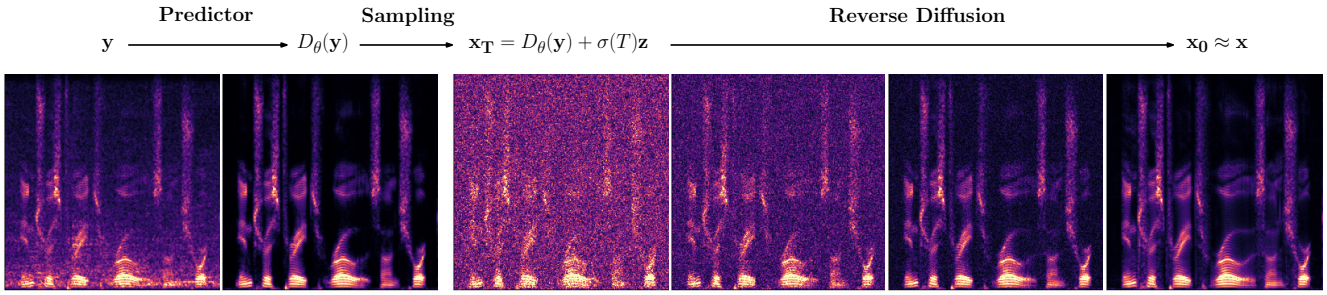
one notices that the resulting *a priori* SNR of the starting point of the reverse process (without accounting for the added Gaussian noise) is very low, as $||\mathbf{x}^{(\text{dis})}||, ||\tilde{\mathbf{n}}|| \ll ||\mathbf{n}||$ for low-enough SNRs and good-enough initial predictor. This makes learning difficult, and we therefore propose to use a different refinement process that we denote as *stochastic regeneration*.

C. Stochastic regeneration

For *stochastic regeneration* we propose to cascade the predictive model D_θ and the generative diffusion model G_ϕ. The generative model learns to *regenerate* the clean speech based on the distorted version provided by the predictive approach. This is conceptually different from the stochastic refinement approach, where the target cues exist in the residual (but are hard to access given the noise present in the noisy residue $\mathbf{r}$) and need to be *refined* by the diffusion model.

The task of the diffusion model is then to guide generation of the clean speech $\mathbf{x}_0$ given the first estimate $D_\theta(\mathbf{y})$. If we look at the decomposition in (11), we simply have to remove the residual noise $\tilde{\mathbf{n}}$ and restore the distorted target cues $\mathbf{x}^{(\mathbf{dis})}$.

Fig. 4: *Proposed stochastic regeneration inference process. The predictive network is first used to generate a denoised version $D_\theta(\mathbf{y})$. Diffusion-based generation G_ϕ is then performed by adding Gaussian noise $\sigma(T)\mathbf{z}$ to obtain the start sample $\mathbf{x}_T$ and solving the reverse diffusion SDE* (10)*, yielding a sample from the estimated posterior $\mathbf{x}_0 \sim p(\mathbf{x}|D_\theta(\mathbf{y}))$.*



The resulting *a priori* SNR in the starting point (again without considering the added Gaussian noise) is very high, as for a reasonable predictor $||\mathbf{x}^{(\text{dis})}||, ||\tilde{\mathbf{n}}|| \ll ||\mathbf{n}||$. The estimate is then obtained as:

$$\hat{\mathbf{x}} = G_\phi(D_\theta(\mathbf{y})) \tag{16}$$

The inference process is shown in Fig. 4. We name the resulting **Sto**chastic **R**egeneration **M**odel *StoRM*.

For training, we use a criterion $\mathcal{J}^{(\text{StoRM})}$ combining denoising score matching and a supervised regularization term —e.g. mean square error—matching the output of the initial predictor to the target speech:

$$\begin{aligned}
\mathcal{J}^{(\text{DSMS})}(\phi) &= \mathbb{E}_{\tau,(\mathbf{x},\mathbf{y}),\mathbf{z}} \left\| \mathbf{s}_\phi(\mathbf{x}_\tau, [\mathbf{y}, D_\theta(\mathbf{y})], \tau) + \frac{\mathbf{z}}{\sigma(\tau)} \right\|_2^2, \\
\mathcal{J}^{(\text{Sup})}(\theta) &= \mathbb{E}_{(\mathbf{x},\mathbf{y})} \|\mathbf{x} - D_\theta(\mathbf{y})\|_2^2, \\
\mathcal{J}^{(\text{StoRM})}(\phi,\theta) &= \mathcal{J}^{(\text{DSMS})}(\phi) + \alpha \mathcal{J}^{(\text{Sup})}(\theta),
\end{aligned} \tag{17}$$

where α is a balance term that we empirically set to 1.

One may object that the estimate $D_\theta(\mathbf{y})$ is not a sufficient statistic for the model to reconstruct target cues. However, by their very learning principle, generative models are able to create data based on the clean examples seen during training, hence our choice of the terminology *stochastic regeneration*. Stochastic regeneration is still a generative model with respect to the definition given in introduction, as it is able to output realistic samples belonging to a posterior distribution. We visually summarize the concepts of predictive, generative and our model in Fig. 5. We describe in algorithms 1 and 2 the training and inference stages of StoRM, respectively.

Algorithm 1 StoRM Training

Input: Training set of pairs $(\mathbf{x}, \mathbf{y})$
Output: Trained parameters $\{\phi, \theta\}$

Sample diffusion time $t \sim \mathcal{U}(t_\epsilon, T)$
Sample noise signal $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$
Infer initial prediction $D_\theta(\mathbf{y})$
Generate perturbed state $\mathbf{x}_\tau \leftarrow \boldsymbol{\mu}(\mathbf{x}, D_\theta(\mathbf{y}), \tau) + \sigma(\tau)\mathbf{z}$
Estimate score $\mathbf{s}_\phi(\mathbf{x}_\tau, [\mathbf{y}, D_\theta(\mathbf{y})], \tau)$
Compute loss $\mathcal{J}^{(\text{StoRM})}(\phi, \theta)$ (eq. (17))
Backpropagate loss $\mathcal{J}^{(\text{StoRM})}(\phi, \theta)$ to update $\{\phi, \theta\}$

Algorithm 2 StoRM Inference (based on PC sampling by [44])

Input: Corrupted speech $\mathbf{y}$, step size $\Delta\tau = \frac{T}{N}$
Output: Clean speech estimate $\widehat{\mathbf{x}}$

Infer initial prediction $D_\theta(\mathbf{y})$
Sample noise signal $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$
Generate initial reverse state $\mathbf{x}_T = D_\theta(\mathbf{y}) + \sigma(T)\mathbf{z}$
for $n \in \{N, \dots, 1\}$ **do**
  Get diffusion time $\tau = n\Delta\tau = \frac{n}{N}T$
  if using corrector **then**
    Estimate score $\mathbf{s}_\phi(\mathbf{x}_\tau, [\mathbf{y}, D_\theta(\mathbf{y})], \tau)$
    Sample correction noise signal $\mathbf{w}_c \sim \mathcal{N}(0, \mathbf{I})$
    Correct estimate (Annealed Langevin Dynamics):
    $\mathbf{x}_\tau \leftarrow \mathbf{x}_\tau + 2r^2\sigma(\tau)^2\mathbf{s}_\phi(\mathbf{x}_\tau, [\mathbf{y}, D_\theta(\mathbf{y})], \tau) + 2r\sigma(\tau)\mathbf{w}_c$
  Estimate score $\mathbf{s}_\phi(\mathbf{x}_\tau, [\mathbf{y}, D_\theta(\mathbf{y})], \tau)$
  Sample prediction noise signal $\mathbf{w}_p \sim \mathcal{N}(0, \mathbf{I})$
  Predict next Euler-Maruyama step:
  $\mathbf{x}_{\tau-\Delta\tau} \leftarrow \mathbf{x}_\tau - \mathbf{s}_\phi(\mathbf{x}_\tau, [\mathbf{y}, D_\theta(\mathbf{y})], \tau)\Delta\tau + \gamma(D_\theta(\mathbf{y}) - \mathbf{x}_\tau)\Delta\tau + g(\tau)\mathbf{w}_p\sqrt{\Delta\tau}$
Output estimate: $\widehat{\mathbf{x}} = \mathbf{x}_0$

IV. Experimental Setup

A. Data

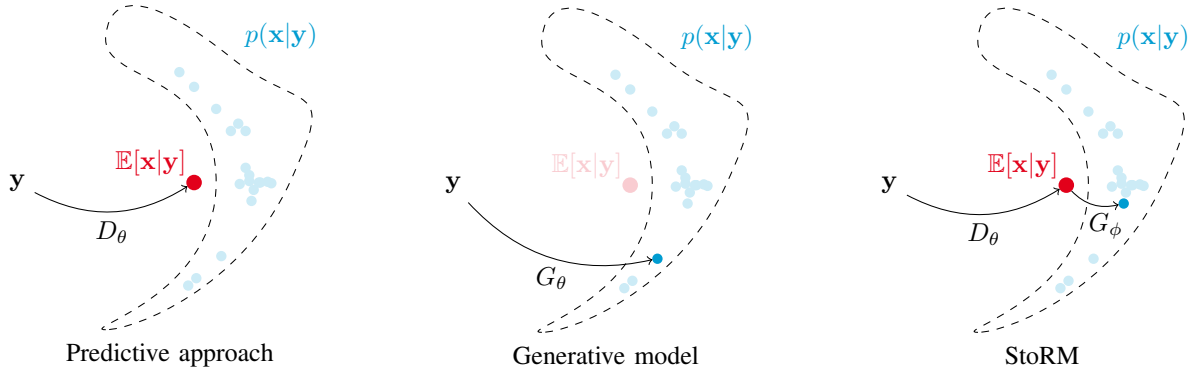
a) Speech Enhancement:

The WSJ0+Chime dataset is generated using clean speech from the WSJ0 corpus [41] and noise signals from the CHiME3 dataset [55]. The mixture signal is created by randomly selecting a noise file and adding it to a clean utterance with a SNR sampled uniformly between -6 and 14 dB.

The TIMIT+Chime dataset is similarly generated as WSJ0+Chime, using TIMIT as the clean speech corpus [42]. We use this dataset for ASR as oracle annotations are available for word error rate (WER) evaluation.

The VoiceBank/DEMAND dataset is a classical benchmark dataset for speech enhancement using clean speech from the VCTK corpus [43] excluding two speakers. The utterances are corrupted by recorded noise from the DEMAND database [56] and two artificial noise types (babble and speech shaped) at

Fig. 5: *Visualization of the inference process for the predictive, generative and proposed StoRM models for a complex posterior distribution (see also Figure 1 in [51]. With the proposed two-stage inference, StoRM uses the predictive mapping to the posterior mean* $\mathbb{E}[\mathbf{x}|\mathbf{y}]$ *as an intermediate step for easier generative inference of a posterior sample* $\mathbf{x}$ *which is more likely to lie in high-density regions of the posterior* $p(\mathbf{x}|\mathbf{y})$



SNRs of 0, 5, 10, and 15 dB for training and validation. The SNR levels of the test set are 2.5, 7.5, 12.5, and 17.5 dB.

b) Speech Dereverberation: The WSJ0+Reverb dataset is generated using clean speech data from the WSJ0 dataset and convolving each utterance with a simulated room impulse response (RIR). We use the `pyroomacoustics` engine [57] to simulate the RIRs. The reverberant room is modeled by sampling uniformly a target T_{60} between 0.4 and 1.0 seconds and a room length, width and height in [5,15]×[5,15]×[2,6] m. This results in an average direct to reverberant ratio (DRR) of around -9 dB and *measured* T_{60} of 0.91 s. A dry version of the room is generated using the same geometric parameters with a fixed absorption coefficient of 0.99, to generate the corresponding anechoic target.

B. Hyperparameters and training configuration

a) Data representation: Utterances are transformed using a short-time Fourier transform (STFT) with a window size of 510, a hop length of 128 and a square-root Hann window, at a sampling rate of 16kHz. A square-root magnitude warping is used to compress the dynamical range of the input spectrograms [12]. For training, sequences of 256 STFT frames (≈2s) are randomly extracted from the full-length utterances and normalized by the maximum absolute value of the noisy utterance before being fed to the network.

b) Forward and reverse diffusion: For all diffusion models, similar values are chosen to parameterize the forward and reverse stochastic processes. The stiffness parameter is fixed to γ =1.5, the extremal noise levels to $\sigma_{\min} = 0.05$ and $\sigma_{\max} = 0.5$, and the extremal diffusion times to $T = 1$ and $\tau_\epsilon = 0.03$ as in [12]. Unless stated otherwise (that is, for all results except those in Figure 6), $N = 50$ time steps are used for reverse diffusion and we adopt the predictor-corrector scheme [44] with one step of annealed Langevin dynamics correction and a step size of $r = 0.5$.

c) Network architecture: The backbone architecture we use is a lighter configuration of the NCSN++ architecture variant proposed in [44], which was used in our previous study [11] and denoted as *NCSN++M*. The following modifications are carried on the up/down-sampling paths of the network: the attention layers are removed (we keep attention in the bottleneck), the number of layers in each encoder-decoder path is decreased from 7 to 4, and only one ResNet block is used per layer instead of two. This results in a network capacity of roughly 27.8M parameters instead of 65M, without significant degradation of the speech enhancement performance, be it for predictive or generative modelling.

When this NCSN++M configuration is used for score estimation in SGMSE+ [12], we call the resulting approach *SGMSE+M*. There, the noisy speech spectrogram $\mathbf{y}$ and the current diffusion process estimate $\mathbf{x}_\tau$ real and imaginary channels are stacked and fed to the network as input, and the current noise level $\sigma(\tau)$ is provided as a conditioner. For our proposed approach StoRM, the initial prediction $D_\theta(\mathbf{y})$ is also stacked together with $\mathbf{y}$ and $\mathbf{x}_\tau$: the influence of this double conditioning is examined in an ablation study in Section V-G. For the predictive approach, denoted directly as *NCSN++M*, the noise-conditioning layers are removed and only the noisy speech spectrogram real and imaginary channels are used. This ablation removes only 1.8% of the original number of parameters, which hardly modifies the network capacity.

We also use ConvTasNet [8] and GaGNet [58] (see next subsection) as alternative initial predictors for StoRM. We train using NCSN++M as the initial predictor and swap it during inference with one of the two networks mentioned above, in order to test the robustness of our proposed stochastic regeneration approach towards unseen predictors (see Section V-D.

d) Baselines: For comparison on WSJ0-based datasets, we compare StoRM to the purely generative SGMSE+M and purely predictive NCSN++M. We also report results using the non-causal version of GaGNet [58], a predictive denoiser using parallel magnitude- and complex-domain processing in the T-F domain. We complement the benchmark on the Voicebank/DEMAND dataset with the non-causal predictive ConvTasNet [8], MetricGAN+ [59], MANNER [60], VoiceFixer [61] as well as the generative unsupervised dynamical VAE (DVAE) [62], conditional time-domain diffusion model CDiffuse [29], stochastic refinement time-domain enhancement scheme SRTNet [37] and original SGMSE [28]. For all

these, we use publicly available code provided by the authors.

e) Training configuration: We use the Adam optimizer [63] with a learning rate of 10^{-4} and an effective batch size of 16. We track an exponential moving average of the DNN weights with a decay of 0.999 to be used for sampling, as it showed to be very effective [64]. We train DNNs for a maximum of 1000 epochs using early stopping based on the validation loss with a patience of 10 epochs. All models converged before reaching the maximum number of epochs. The generative approach is trained with the denoising score matching criterion (8), and the predictive methods use a simple mean-square error loss on the complex spectrogram. The stochastic regeneration approach uses the combined criterion in (17). The default training strategy is that we pre-train the initial predictor with a simple mean-square error loss, then jointly train the predictor and score networks with (17). Different training strategies are examined in an ablation study in Section V-G.

C. Evaluation metrics

For instrumental evaluation of the speech enhancement and dereverberation performance with clean test data available, we use intrusive measures such as Perceptual Evaluation of Speech Quality (PESQ) [65] to assess speech quality, extended short-term objective intelligibility (ESTOI) [66] for intelligibility and scale-invariant signal to distortion ratio (SI-SDR), scale-invariant signal to interference ratio (SI-SIR) and scale-invariant signal to artifacts ratio (SI-SAR) [67] for noise removal. As in [11], we complement our metrics benchmark with WV-MOS [68] , which is a DNN-based mean opinion score (MOS) estimation, and was used by the authors for reference-free assessment of bandwidth extension or speech enhancement performance.

We also evaluate our proposed approach on ASR, using NVidia's temporal convolutional network QuartzNet [69] as the speech recognition model, and classical WER dynamic programming evaluation with the `jiwer` Python library[2]. We use the pretrained `Base-en` 18.9M parameters version of QuartzNet for specialized English speech recognition.

Finally, we organize a medium-scale MUSHRA listening test with 9 participants. We ask the participants to rate 10 samples with a single number representing overall quality, including speech distortion, residual distortions and potential artifacts. We use the `webMUSHRA`[3] tool with `pymushra`[4] server management. The samples are randomly extracted from the WSJ0+Chime and WSJ0+Reverb test sets, ensuring gender and task balance as well as speaker exclusivity (within a given task, a speaker is used once at most). The approaches evaluated are the predictive NCSN++M, score-based generative model SGMSE+ and our proposed approach StoRM. The noisy mixture is given as a low anchor, and a supplementary anchor is created by increasing the input SNR by $10\mathrm{dB}$ in comparison to the noisy mixture.

[2]https://github.com/jitsi/jiwer
[3]https://github.com/audiolabs/webMUSHRA
[4]https://github.com/nils-werner/pymushra

V. Experimental Results and Discussion

A. Comparison to baselines

a) WSJ0+Chime and WSJ0+Reverb: In tables I and II, we show results of the proposed stochastic regeneration StoRM approach as compared to purely predictive GaGNet and NCSN++M and purely generative SGMSE+M, for denoising on WSJ0+Chime and dereverberation on WSJ0+Reverb. Based on preliminary experiments we confirm that the approach can be successfully trained for joint dereverberation and denoising, and we refer the reader to our web page where audio samples are presented for this joint task. In this work, however, we separate the two tasks to get more insights into the individual performance.

We confirm the results from [11], which is that predictive NCSN++M and GaGNet provide samples with good interference removal (high SI-SDR) and intelligibility (high ESTOI) but lower quality (lower PESQ and WV-MOS) compared to diffusion-based generative SGMSE+M. This gap is stronger for dereverberation than for denoising as already observed, since the average input SNR for dereverberation is much lower than for denoising. Also, the reverberation interference being a filtered version of the target speech, the predictive method cannot suppress reverberation without introducing significant distortion, which is particularly audible in NCSN++M and GaGNet results. The generative SGMSE+M, however, is able to extract the speech cues and directly reconstructs with hardly any reverberation left.

It is generally observed that point-wise measures like SI-SDR, SI-SIR and SI-SAR provide generally worse results for generative models than for predictive models [6], [24]. This is because generative models try to estimate the posterior distribution, providing better perceptual metrics, whereas predictive models are implicitly trained to recover the posterior mean and average out the distortions for each point, thus yielding higher point-wise fidelity [6], [11].

We observe that our proposed StoRM associates the best of both the predictive and generative worlds, by producing samples with very high quality like generative SGMSE+M, while being approximately as good with interference removal as the predictive NCSN++M. Again, the observed gap is more significant for dereverberation, where the proposed StoRM outperforms both SGMSE+M and NCSN++M on all metrics. Example spectrograms are displayed on Figure 7 and 8, for denoising and dereverberation respectively.

b) VoiceBank/DEMAND: We report in Table III results of our StoRM configuration against various state-of-the-art speech enhancement baselines on the VoiceBank/DEMAND benchmark. The SNRs in Voicebank/DEMAND are always positive and distributed around $10\mathrm{dB}$, which is not very challenging compared to the conditions in our WSJ0+Chime dataset. Consequently, the gap between SGMSE+ and StoRM on Voicebank/DEMAND is not as large as on WSJ0+Chime, which shows that using the initial predictor is particularly useful in difficult conditions. In easier environments such as that simulated in Voicebank/DEMAND, diffusion-based generative modelling can take the noisy mixture as the initial condition for reverse diffusion without being further guided.

TABLE I: *Denoising results obtained on WSJ0+Chime. Values indicate mean and standard deviation. All approaches (except GaGNet) use the NCSN++M architecture. Diffusion models (SGMSE+M and StoRM) use $N = 50$ steps for reverse diffusion.*

Method	WV-MOS	PESQ	ESTOI	SI-SDR	SI-SIR	SI-SAR
Mixture	1.43 ± 0.66	1.38 ± 0.32	0.65 ± 0.18	4.3 ± 5.8	4.3 ± 5.8	-
SGMSE+M	3.63 ± 0.38	2.33 ± 0.61	0.86 ± 0.10	13.3 ± 5.0	27.4 ± 6.3	13.5 ± 4.9
NCSN++M	3.47 ± 0.53	2.21 ± 0.65	**0.89 ± 0.09**	**16.4 ± 4.4**	31.1 ± 5.0	**16.6 ± 4.4**
GaGNet	3.34 ± 0.54	2.19 ± 0.61	0.87 ± 0.09	15.7 ± 4.3	27.6 ± 4.7	16.0 ± 4.4
StoRM	**3.72 ± 0.40**	**2.58 ± 0.61**	0.88 ± 0.08	15.1 ± 4.2	**31.6 ± 5.0**	15.3 ± 4.2

TABLE II: *Dereverberation results on Reverb-WSJ0. Values indicate mean and standard deviation. All approaches (except GaGNet) use the NCSN++M architecture. Diffusion models (SGMSE+M and StoRM) use $N = 50$ steps for reverse diffusion.*

Method	WV-MOS	PESQ	ESTOI	SI-SDR	SI-SIR	SI-SAR
Mixture	1.78 ± 0.99	1.36 ± 0.19	0.46 ± 0.12	-7.3 ± 5.5	-7.5 ± 5.4	-
SGMSE+M	3.49 ± 0.39	2.66 ± 0.45	0.85 ± 0.06	2.4 ± 7.2	11.6 ± 9.9	2.8 ± 6.8
NCSN++M	2.99 ± 0.38	2.08 ± 0.47	0.85 ± 0.06	6.1 ± 3.8	21.4 ± 7.0	6.1 ± 3.7
GaGNet	2.40 ± 0.52	1.59 ± 0.37	0.68 ± 0.09	-0.5 ± 4.8	7.7 ± 4.0	0.2 ± 5.1
StoRM	**3.73 ± 0.32**	**2.83 ± 0.42**	**0.88 ± 0.04**	**6.5 ± 4.0**	**22.9 ± 8.2**	**6.5 ± 3.9**

TABLE III: *Denoising results obtained on VoiceBank/DEMAND.* P *means predictive and* G *generative. All approaches were evaluated on the test set using publicly available code attached to the respective papers.*

Method	Type	PESQ	ESTOI	SI-SDR	WV-MOS
Mixture		1.97	0.79	8.4	2.99
NCSN++	P	2.83	**0.88**	**20.1**	4.07
NCSN++M	P	2.82	**0.88**	19.9	4.06
Conv-TasNet [8]	P	2.84	0.85	19.1	4.28
MetricGAN+ [59]	P	3.13	0.83	8.5	3.90
MANNER [60]	P	**3.21**	0.87	18.9	**4.38**
GaGNet [58]	P	2.94	0.86	18.1	4.23
VoiceFixer [61]	P	2.11	0.73	-4.3	4.00
DVAE [62]	G	2.43	0.81	16.4	3.73
CDiffuSE [29]	G	2.46	0.79	12.6	3.64
SRTNet [37]	G	2.11	0.81	8.5	3.58
SGMSE [28]	G	2.28	0.80	16.2	3.90
SGMSE+ [12]	G	2.93	0.87	17.3	4.24
SGMSE+M	G	**2.96**	0.87	17.3	4.26
StoRM (proposed)	G	2.93	**0.88**	**18.8**	**4.30**

Still, our proposed method StoRM still slightly outperforms the other generative models on ESTOI, WV-MOS and SI-SDR, setting a new state-of-the-art record for generative models on this benchmark.

B. Efficient sampling

We report in Figure 6 the performance of the SGMSE+M and StoRM schemes as a function of the number of steps used for reverse diffusion. We additionally provide an estimation of the number of multiply-accumulate (MAC) operations per second as measured by the `python-papi` package.

We observe that StoRM is able to maintain performance at a near-optimal level even using only 10 steps, using the initial predictive estimate as a reasonable guess for further diffusion. In comparison, SGMSE+M performance degrades rapidly as the number of steps decreases. Furthermore, StoRM is able to produce very high-quality samples without even needing the Annealed Langevin Dynamics corrector during sampling, whereas SGMSE+M performance dramatically degrades without this corrector. Since each corrector step makes an additional call to the score network, avoiding its use further relaxes the computational complexity. StoRM therefore highlights a strong compromise between inference speed and sample quality. Using StoRM with 20 steps and no corrector produces near-optimal sample quality at a cost of 4.5 $\cdot 10^{11}$ MAC$\cdot$s^{-1}, versus 2.1 $\cdot 10^{12}$ MAC$\cdot$s^{-1} for the optimal SGMSE+M setting (50 steps and Annealed Langevin Dynamics correction). StoRM even outperforms the optimal SGMSE+M setting using 10 steps and no corrector, thus reducing computational complexity by a full order of magnitude. In our recent work [70], we proposed to make our diffusion generative model causal to demonstrate its application to real-world scenarios.

C. Generalization to unseen data

In Table IV, we examine robustness to mismatched training and test data. The mismatched condition is generated by training on Voicebank/DEMAND and testing on WSJ0+Chime*. WSJ0+Chime* is created in the same fashion as WSJ0+Chime, but the input SNRs range between 0 and 20dB to match the SNR range in Voicebank+DEMAND, and thus constitutes the same benchmark as in [12]. We observe that NCSN++M shows a reasonable generalization ability because of its sophisticated network architecture. However, SGMSE+M and StoRM's ability to maintain performance as compared to the matched case is even superior. This shows that StoRM can leverage generative modelling to correct the relative lack of

Fig. 6: *Results for denoising on WSJ0+Chime as a function of the number of reverse diffusion steps N. All approaches use the same NCSN++M architecture. The corrector uses one step of Annealed Langevin Dynamics with $r = 0.5$.*

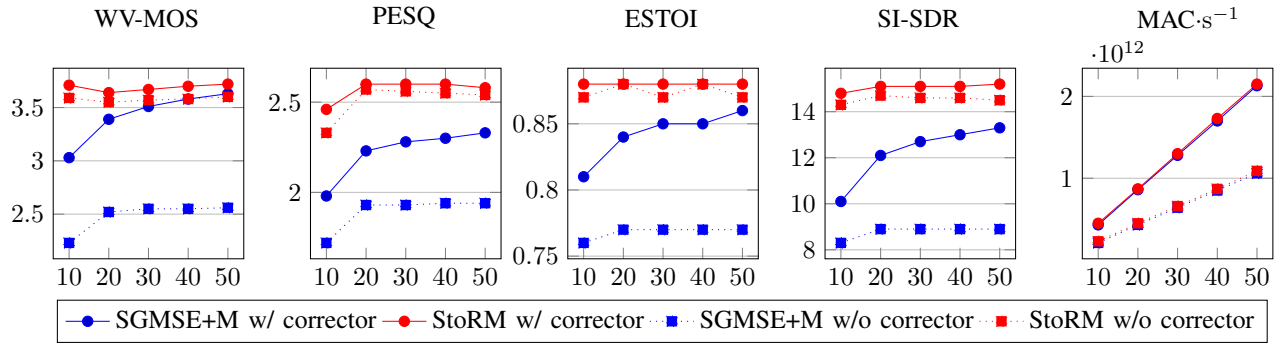


Fig. 7: *Clean, noisy and processed utterances from WS0+Chime. Input SNR is -0.9* dB. *Vocalizing artifacts are visible at the beginning of the SGMSE+M utterance. Speech distortions are observed in the NCSN++M sample. StoRM corrects these distortions without introducing vocalizing artifacts and yield high quality and intelligibility.*

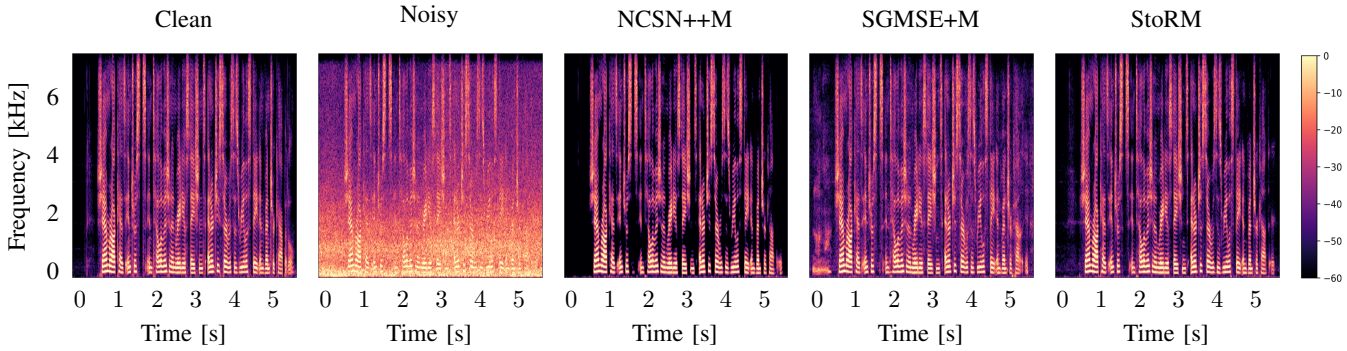


Fig. 8: *Anechoic, reverberant and processed utterances from WS0+Reverb. Input T_{60} is 1.06* s. *Formant structure is partly destroyed by SGMSE+M and severe speech distortions are observed in the NCSN++M sample. StoRM corrects the distortions and reproduces the formant structure without residual reverberation.*

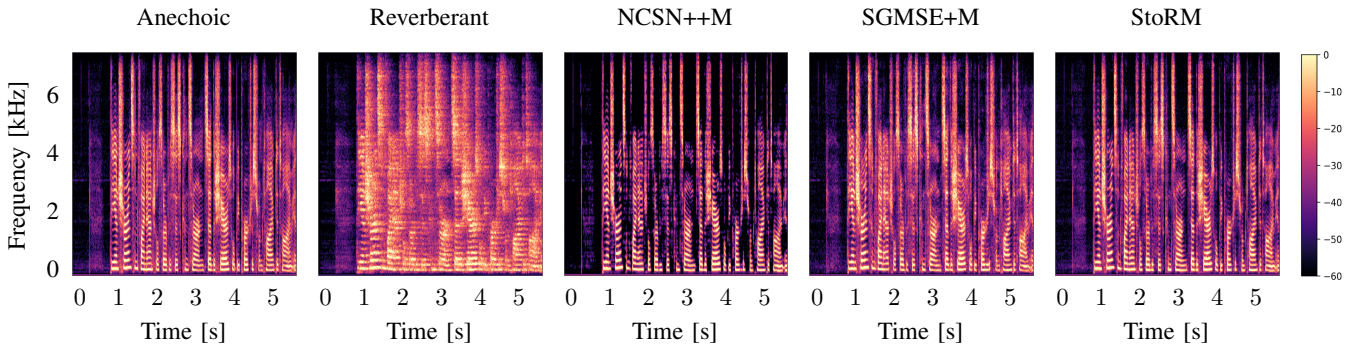


TABLE IV: *Denoising results on WSJ0+Chime⋆ (with input SNRs between 0dB and 20dB) in matched and mismatched settings. In the mismatch setting, the approaches are trained on VoiceBank/DEMAND. All methods use the NCSN++M architecture. SGMSE+M and StoRM use* 50 *diffusion steps.*

Method	Match	WV-MOS	PESQ	ESTOI	SI-SDR
NCSN++M	✓	3.78	2.73	**0.94**	**20.0**
	✗	3.33 (-0.35)	2.14 (-0.59)	**0.90** (-0.04)	**17.6** (-2.4)
SGMSE+M	✓	3.84	2.96	0.92	17.4
	✗	3.61 (-0.23)	**2.48** (**-0.48**)	**0.90** (**-0.02**)	15.7 (**-1.7**)
StoRM	✓	**3.93**	**3.01**	0.93	18.5
	✗	**3.71** (**-0.22**)	2.47 (-0.54)	**0.90** (-0.03)	16.8 (**-1.7**)

robustness of the first predictive stage to this mismatched condition.

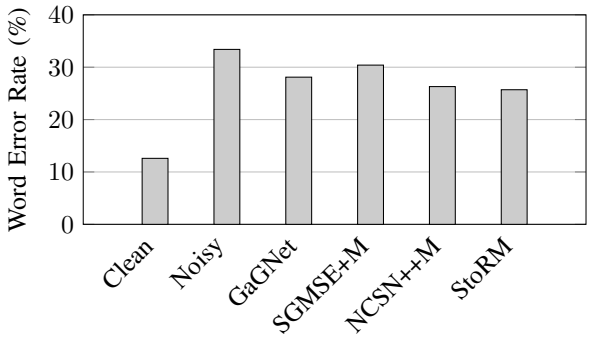
D. Generalization to mismatched predictors

In Table V, we report results for StoRM using different initial predictors than the one used during training. The approach is trained using the NCSN++M as initial predictor as before, and we test using ConvTasNet [8] and GaGNet [58] as alternative initial predictors by exchanging this predictor during inference. We observe that the artifacts in GaGNet estimates are of a similar nature than those of NCSN++M, as both approaches process speech in the T-F domain. StoRM

TABLE V: *Denoising results on WSJ0+Chime for StoRM using matched and mismatched initial predictors. The predictor architecture used for training is NCSN++M. All approaches use the NCSN++M as score network and $N = 50$ steps . Values indicate mean and standard deviation.*

Initial Predictor	Matched	PESQ	ESTOI	SI-SDR
Mixture	-	1.38 ± 0.32	0.65 ± 0.18	4.3 ± 5.8
NCSN++M	✓	**2.53 ± 0.63**	**0.88 ± 0.09**	**14.7 ± 4.3**
GaGNet [58]	✗	2.52 ± 0.62	0.87 ± 0.09	**14.7 ± 4.1**
ConvTasNet [8]	✗	2.36 ± 0.60	0.86 ± 0.09	9.9 ± 1.7

Fig. 9: *ASR results for speech enhancement on TIMIT+Chime. Diffusion models use $N = 50$ steps.*



is entirely robust to such a slight mismatch, as indicated by the equivalent performance of using NCSN++M and GaGNet as the initial predictor. ConvTasNet is a time-domain method using a fully learnt encoder: the speech distortions are then different than those of NCSN++M or GaGNet. Additionally, ConvTasNet's original performance is slightly worse than its two counterparts. Consequently, we observe that the performance of StoRM using ConvTasNet as the initial predictor is poorer but close to that of using NCSN++M as the predictor. This demonstrates relative robustness to unseen conditions provided by the generative modelling stage.

E. ASR results

In Figure 9, we compare the predictive, generative and stochastic regeneration approaches on speech enhancement for ASR using the TIMIT+Chime dataset. We observe that SGMSE+M results in poorer speech recognition abilities than its predictive GaGNet and NCSN++M counterparts, and hardly improves the ASR performance over the noisy mixture. This can be explained by the previously mentioned undesired vocalizing artifacts and phonetic confusions, which are created by the generative approach under uncertainty over the presence and phonetic nature of speech respectively and are heavily punished by WER evaluation. Using the predictive estimate as a guide for generation, StoRM improves the WER performance by a relative factor of 19% as compared to SGMSE+M, even slightly outperforming NCSN++M, which shows that most of the artifacts and confusions are corrected.

Fig. 10: *Listening test results. CQS is the "continuous quality scale" on which participants are asked to rate. Inner line represents the median. 9 participants rated 10 samples randomly selected from WSJ0+Chime and WSJ0+Reverb.*

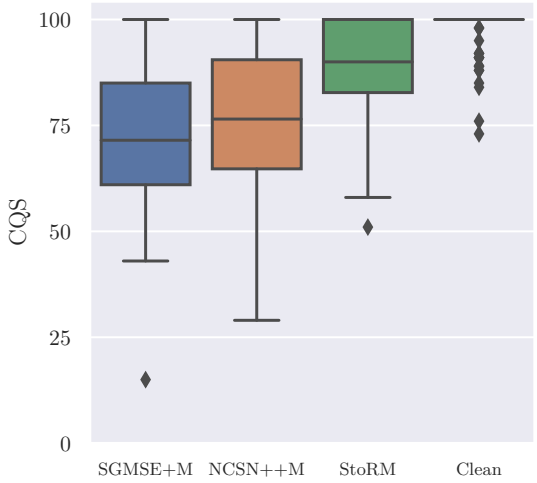


TABLE VI: *Denoising results on WSJ0+Chime for StoRM using different conditioning inputs for the score network. Values indicate mean and standard devation. All approaches use NCSN++M as backbone architecture and $N = 50$ steps.*

Conditioning	PESQ	ESTOI	SI-SDR
Noisy	2.30 ± 0.60	0.84 ± 0.10	11.5 ± 5.2
PostDenoiser	2.50 ± 0.62	0.87 ± 0.09	14.7 ± 4.3
Both	**2.53 ± 0.63**	**0.88 ± 0.08**	**15.1 ± 4.2**

F. Listening experiment

We show in the boxplot on Figure 10 the results of our MUSHRA listening test. On average, the participants clearly rated the proposed StoRM higher than the purely predictive NCSN++M and purely generative SGMSE+M. This confirms the results provided by the intrusive and non-intrusive metrics provided in tables I and II. Participants rated NCSN++M slightly better than SGMSE+M on average, which is linked to the rating criterion described in Section IV-C. This seems to indicate that participants put more weight on "residual distortions" and "potential artifacts" (vocalizing/breathing/confusions) than on "speech distortion".

G. Ablation studies

We conduct ablation studies on the WSJ0+Chime dataset, to observe the respective influence of score network conditioning on the one hand and the training strategy on the other hand.

a) Conditioning of the score network: In Table VI, we report instrumental results when using different conditioning inputs for the score network used in the proposed StoRM. We input either the noisy speech $\mathbf{y}$ ("Noisy"), the denoised estimate $D_\theta(\mathbf{y})$ ("PostDenoiser"), or both ("Both", which is the default setting for StoRM). Using only the noisy speech ("Noisy") is detrimental to the performance. It seems that the

TABLE VII: *Denoising results on WSJ0+Chime for StoRM using different training strategies for the score network. All approaches use NCSN++M as backbone architecture and $N = 50$ steps for reverse diffusion.*

Pre-train D_θ	Fine-tune D_θ	Use $\mathcal{J}^{(\mathrm{Sup})}$	PESQ	ESTOI	SI-SDR
✗	✓	✓	**2.58**	**0.88**	**15.1**
✓	✗	✗	2.53	**0.88**	14.7
✗	✓	✗	1.11	0.62	-0.3
✓	✓	✓	**2.58**	**0.88**	**15.1**

score network does need the information from the original distortions in $D_\theta(\mathbf{y})$ at time step τ=T, to properly learn the score at time step τ<T. This mismatch at the first denoising steps is detrimental to performance. We also observe that instrumental metrics tend to slightly favor the "Both" conditioning over the "PostDenoiser" conditioning.

b) Training strategies: We show in Table VII the results of StoRM using different training strategies. We see that jointly training the initial predictor and the score network slightly improves results for denoising. However, training the initial predictor from scratch or having it pre-trained first does not seem to make a difference, as long as one regularizes the training criterion with the supervised criterion $\mathcal{J}^{(\mathrm{Sup})}$ which matches the output of the initial predictor to the target. Indeed, as shown in the third line of Table VII, if we use a randomly initialized predictor and train both the predictor and score networks only with the score matching criterion $\mathcal{J}^{(\mathrm{DSM})}$—i.e. setting α in (17) to 0—the performance dramatically drops. This is to be expected since the learning task then becomes much more complicated given the size of the search space and the lack of regularization. The proposed combination of joint training and regularization with $\mathcal{J}^{(\mathrm{Sup})}$ performs most favorably. This indicates that it is best to train the predictor to output something resembling clean speech rather than arbitrary learned encoder features, while still leaving some room for the predictor to adapt its output to the score model. Our experiments also indicate that, once pre-trained, StoRM converges 25% faster than when training from scratch. However, the cumulated time of pre-training the predictor and fine-tuning is 50% larger than the cost of training from scratch.

VI. Conclusion

We presented a generative stochastic regeneration scheme combining a predictive model as initial predictor and a diffusion-based generative approach regenerating the target cues distorted by the first stage. On the one hand, the approach improves sample quality compared to pure predictive approaches as it leverages generative modelling to output samples that have high probability on the target posterior distribution manifold, rather than regressing to their mean. On the other hand, it uses predictive power to provide a good initial prediction of the target sample, which avoids typical generative artifacts such as vocalizing and breathing effects, and increases the interference removal performance, especially in difficult environments. Intrusive and reference-free instrumental metrics as well as formal listening tests confirmed the superiority of the stochastic regeneration approach over the baselines. The resulting approach allows efficient sampling, requiring fewer steps and avoiding the use of Annealed Langevin Dynamics correction during reverse diffusion, thus reducing computational complexity by an order of magnitude without sacrificing quality, compared to the original diffusion model.

References

[1] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Noise Control Engineering Journal, 2011, vol. 59.

[2] S. J. Godsill, P. J. W. Rayner, and O. Cappé, *Digital audio restoration*. Springer, 1998.

[3] 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, 2013.

[4] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement*, E. Vincent, Ed. John Wiley & Sons, 2018.

[5] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. Audio, Speech, Language Proc.*, 2018.

[6] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[7] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 25, no. 7, pp. 1492–1501, 2017.

[8] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network," in *ISCA Interspeech*, 2018.

[9] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen, "A two-stage approach to speech bandwidth extension," in *ISCA Interspeech*, 2021.

[10] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 23, no. 6, pp. 982–992, 2015.

[11] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2023.

[12] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE Trans. Audio, Speech, Language Proc.*, 2023.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *Int. Conf. Learning Repr. (ICLR)*, 2014.

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

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

[16] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks." *ISCA Interspeech*, 2020.

[17] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Int. Conf. Machine Learning (ICML)*, 2015.

[18] I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3964–3979, nov 2021.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Neural Information Processing Systems (NIPS)*, vol. 27, 2014.

[20] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Neural Information Processing Systems (NIPS)*, 2020.

[21] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *Int. Conf. Machine Learning (ICML)*, 2015.

[22] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Neural Information Processing Systems (NIPS)*, 2020.

[23] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Neural Information Processing Systems (NIPS)*, vol. 32, 2019.

[24] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar, "Deblurring via stochastic refinement," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[25] S. Welker, H. N. Chapman, and T. Gerkmann, "DriftRec: Adapting diffusion models to blind image restoration tasks," *arXiv preprint*, 2022.

[26] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Neural Information Processing Systems (NIPS)*, vol. 34, 2021.

[27] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *arXiv preprint*, 2022.

[28] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *ISCA Interspeech*, 2022.

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

[30] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv preprint*, 2022.

[31] S. Han and J. Lee, "Nu-wave 2: A general neural audio upsampling model for various sampling rates," in *ISCA Interspeech*, 2022.

[32] D. P. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," in *Neural Information Processing Systems (NIPS)*, 2021.

[33] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Int. Conf. Learning Repr. (ICLR)*, 2021.

[34] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "BDDM: Bilateral denoising diffusion models for fast and high-quality speech synthesis," in *Int. Conf. Learning Repr. (ICLR)*, 2022.

[35] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[36] B. Jing, G. Corso, R. Berlinghieri, and T. Jaakkola, "Subspace diffusion generative models," in *European Conf. on Computer Vision (ECVA)*, 2022.

[37] Z. Qiu, M. Fu, Y. Yu, L. Yin, F. Sun, and H. Huang, "SRTNet: Time domain speech enhancement via stochastic refinement," *arXiv preprint*, 2022.

[38] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Neural Information Processing Systems (NIPS)*, 2022.

[39] K. Saito, N. Murata, T. Uesaka, C.-H. Lai, Y. Takida, T. Fukui, and Y. Mitsufuji, "Unsupervised vocal dereverberation with diffusion-based generative models," *arXiv preprint*, 2022.

[40] R. Sawata, N. Murata, Y. Takida, T. Uesaka, T. Shibuya, S. Takahashi, and Y. Mitsufuji, "A versatile diffusion-based generative refiner for speech enhancement," *arXiv preprint*, 2022.

[41] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete." [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S6A

[42] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.

[43] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," *9th ISCA Speech Synthesis Workshop*, 2016.

[44] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *Int. Conf. Learning Repr. (ICLR)*, 2021.

[45] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching." *Journal of Machine Learning Research*, 2005.

[46] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[47] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*. Journal of the American Statistical Association, 2000, vol. 82.

[48] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, 1982.

[49] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," *Int. Conf. Learning Repr. (ICLR)*, 2021.

[50] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.

[51] M. Delbracio and P. Milanfar, "Inversion by direct iteration: An alternative to denoising diffusion for image restoration," *arXiv preprint*, 2023.

[52] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Neural Information Processing Systems (NIPS)*, 2017.

[53] A. Avila, A. Alam, D. O'Shaughnessy, and T. Falk, "Investigating speech enhancement and perceptual quality for speech emotion recognition," in *ISCA Interspeech*, 2018.

[54] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.

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

[56] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[57] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Apr. 2018.

[58] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, 2022.

[59] 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 *ISCA Interspeech*, 2022, pp. 7412–7416.

[60] H. J. Park, B. H. Kang, W. Shin, J. S. Kim, and S. W. Han, "Manner: Multi-view attention network for noise erasure," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2022.

[61] H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "Voicefixer: Toward general speech restoration with neural vocoder," *arXiv preprint*, 2021.

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

[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int. Conf. Learning Repr. (ICLR)*, 2015.

[64] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *Neural Information Processing Systems (NIPS)*, 2020.

[65] 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)*, 2001.

[66] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, 2016.

[67] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr - half-baked or well done?" in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2019.

[68] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: a unified framework for bandwidth extension and speech enhancement," *arXiv preprint*, 2022.

[69] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook *et al.*, "Nemo: a toolkit for building AI applications using neural modules," *arXiv preprint*, 2019.

[70] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Speech signal improvement using causal generative diffusion models," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2023.

5

Solving Single-Channel Speech Dereverberation as an Inverse Problem with Unsupervised Diffusion Models

5.1 Unsupervised Diffusion Models for Informed Dereverberation [P10]

Abstract

We present in this paper an informed single-channel dereverberation method based on conditional generation with diffusion models. With knowledge of the room impulse response, the anechoic utterance is generated via reverse diffusion using a measurement consistency criterion coupled with a neural network that represents the clean speech prior. The proposed approach is largely more robust to measurement noise compared to a state-of-the-art informed single-channel dereverberation method, especially for non-stationary noise. Furthermore, we compare to other blind dereverberation methods using diffusion models and show superiority of the proposed approach for large reverberation times. We motivate our algorithm by introducing an extension for blind dereverberation allowing joint estimation of the room impulse response and anechoic speech. Audio samples and code can be found online.

Reference

Jean-Marie Lemercier, Simon Welker and Timo Gerkmann, "Diffusion Posterior Sampling for Informed Single-Channel Dereverberation", *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, 2023, DOI: 10.1109/WASPAA58266.2023.10248108

Copyright Notice

Authors' Contributions

Jean-Marie Lemercier is the first author of this publication. He implemented all algorithms, trained the neural networks used in the paper, conducted the experimental validation, and wrote the manuscript. Simon Welker brought some feedback through discussions on diffusion posterior sampling schemes and also helped with reviewing the manuscript. Timo Gerkmann brought insights on the experimental validation, mathematical derivations, and reviewed the manuscript.

DIFFUSION POSTERIOR SAMPLING FOR INFORMED SINGLE-CHANNEL DEREVERBERATION

Jean-Marie Lemercier[1*], *Simon Welker*[1,2†], *Timo Gerkmann*[1]

[1] Signal Processing (SP), Universität Hamburg, Germany
[2] Center for Free-Electron Laser Science, DESY, Hamburg, Germany

ABSTRACT

We present in this paper an informed single-channel dereverberation method based on conditional generation with diffusion models. With knowledge of the room impulse response, the anechoic utterance is generated via reverse diffusion using a measurement consistency criterion coupled with a neural network that represents the clean speech prior. The proposed approach is largely more robust to measurement noise compared to a state-of-the-art informed single-channel dereverberation method, especially for non-stationary noise. Furthermore, we compare to other blind dereverberation methods using diffusion models and show superiority of the proposed approach for large reverberation times. We motivate our algorithm by introducing an extension for blind dereverberation allowing joint estimation of the room impulse response and anechoic speech. Audio samples and code can be found online[1].

Index Terms— Informed dereverberation, diffusion models, posterior sampling, inverse problems

1. INTRODUCTION

Reverberation is a natural phenomenon occurring in most spaces of our daily life, where sound waves get reflected and attenuated by the enclosure walls. It degrades speech intelligibility and quality for normal listeners, and dramatically so for hearing-impaired listeners [1]. Therefore, modern communication devices and listening setups are equipped with dereverberation algorithms which aim to recover the anechoic component of speech [1]. We will denote as *informed* the methods that exploit prior knowledge of the room impulse response (RIR) and as *blind* the methods that try to recover anechoic speech without knowing the RIR.

Traditional blind dereverberation methods exploit the statistical properties of the anechoic and reverberant signals, typically in the time, spectral or cepstral domain [2]. Machine learning techniques try to learn these statistical properties directly from data [3]. Typically, supervised predictive models for blind dereverberation include time-frequency (T-F) maskers [4], time domain methods [5] and direct spectro-temporal mapping [6]. Generative models, that aim to learn the posterior distribution of clean speech conditioned on corrupted speech, have also been introduced for blind dereverberation and speech enhancement. In particular, conditional diffusion-based generative models (or simply *diffusion models*) [7, 8] have been successfully applied to blind dereverberation [9–11].

*Funded by the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380. Corresponding author: jeanmarie.lemercier@uni-hamburg.de

†Funded by DASHH (Data Science in Hamburg Hemholtz Graduate School for the Structure of Matter) with the Grant No. HIDSS-0002.

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

Though informed dereverberation may seem an easier task in comparison to blind dereverberation, knowing the RIR does not guarantee to find a stable and causal inverse filter in the single-channel case, as typical real-world RIRs are mixed-phase signals [12]. Using multiple microphones may mend such issues to some extent [13], but may also suffer from limited robustness [14]. Single-channel informed methods include least-squares and $\mathcal{L}^p$-based optimization rules [15–17], frequency-domain methods such as homomorphic inverse filtering [15] , and hybrid techniques such as [18] where a regularized inverse filter is used to avoid non-causality artifacts and a speech enhancement scheme is used as a post-processing step to attenuate residual pre-echoes.

In this paper, we present a single-channel informed dereverberation technique using diffusion models, with two variants for reverse sampling. We show that the proposed method retrieves high-quality anechoic speech samples for all reverberant conditions without the need for post-processing. We also demonstrate the robustness of the proposed method to measurement noise. We compare our results with a state-of-the-art frequency-domain informed dereverberation method [18] as well as recently introduced diffusion models for blind dereverberation [9,10]. Code and audio examples are provided in the supplementary material.

2. DIFFUSION-BASED GENERATIVE MODELS

In this section we introduce diffusion models, a class of generative models that has recently showed impressive abilities to learn natural data distributions in the image [7, 8] and speech domains [9, 11, 19]. Score-based diffusion models in the framework by Song et al. [8] are defined by three components: a forward diffusion process parameterized by a stochastic differential equation (SDE), a score estimator implemented by a deep neural network (DNN) and a sampling method for inference.

As in [9,10,19,20], here the processes are defined in the complex spectrogram domain, independently for each T-F bin. In the following, the variables in uppercase bold are assumed to be vectors $\mathbb{C}^D$ containing coefficients of a flattened complex spectrogram— with D the product of the time and frequency dimensions— whereas variables in lowercase bold are time vectors in $\mathbb{R}^L$ (unless specified) and variables in regular font are scalars in $\mathbb{C}$. The stochastic *forward process* $\{\mathbf{X}_\tau\}_{\tau=0}^{T}$ slowly transforms clean speech into a tractable noise distribution . It is modeled as the solution to the following Variance-Exploding SDE [8]:

$$\mathrm{d}\mathbf{X}_\tau = g(\tau)\mathrm{d}\mathbf{W}_\tau, \tag{1}$$

$$g(\tau) = \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\tau}\sqrt{2\log\frac{\sigma_{\max}}{\sigma_{\min}}}, \tag{2}$$

where $\mathbf{X}_\tau$ is the current state of the process indexed by a continu-

ous time variable $\tau \in [0, T]$. The stochastic process $\mathbf{W}_\tau$ is a standard D-dimensional Brownian motion, which implies that $\mathrm{d}\mathbf{W}_\tau$ is a zero-mean Gaussian random variable with infinitesimal standard deviation for each T-F bin. The initial condition $\mathbf{X}_0 = \mathbf{X}$ represents clean speech and the *diffusion coefficient* g controls the amount of white noise injected at each step, with $\sigma_{\min}$ and $\sigma_{\max}$ being hyperparameters representing extremal noise levels.

The *reverse process* $\{\mathbf{X}_\tau\}_{\tau=T}^{0}$ turning noise into clean speech is another diffusion process also defined as the solution of a SDE [8, 21], with τ flowing in reverse (i.e. $\mathrm{d}\tau < 0$). Here, we will use the corresponding *probability flow* ordinary differential equation (ODE), since its solution has the same marginal distribution as its SDE counterpart [8]:

$$\mathrm{d}\mathbf{X}_\tau = -\frac{1}{2} g(\tau)^2 \nabla_{\mathbf{X}_\tau} \log p(\mathbf{X}_\tau) \mathrm{d}\tau. \tag{3}$$

The quantity $\nabla_{\mathbf{X}_\tau} \log p(\mathbf{X}_\tau)$ is the *score function*, i.e. the gradient of the logarithm distribution for the current state $\mathbf{X}_\tau$. At inference time, this score function is not available, and therefore a neural network $s_\theta(\mathbf{X}_\tau, \sigma(\tau))$, called *score model*, is used to estimate the score of the current state $\mathbf{X}_\tau$ given the current Gaussian noise standard deviation $\sigma(\tau)$. The latter encodes how much Gaussian noise is left to remove before getting in the vicinity of clean speech $\mathbf{X}_0$. It must therefore be fed to the score network as conditioning, and is obtained in closed-form for the Variance-Exploding SDE [8]. The score model s_θ is trained via *denoising score matching* [22].

3. DIFFUSION POSTERIOR SAMPLING FOR DEREVERBERATION

3.1. Diffusion Posterior Sampling for Inverse Problems

Inverse problems consist in finding the state $\mathbf{X}$ given a observation $\mathbf{Y} = \mathcal{A}(\mathbf{X})$ with $\mathcal{A}$ being a measurement operator. We consider the *non-blind noisy linear* inverse problem of informed single-channel dereverberation. That is, we wish to retrieve the anechoic version of some reverberant speech under measurement noise $\mathbf{n}$ when the RIR $\mathbf{k} \in \mathbb{R}^K$ is known. We define the mixing process in the time-domain, with $\mathbf{x} := \mathrm{iSTFT}(\mathbf{X})$ and $\mathbf{y} := \mathrm{iSTFT}(\mathbf{Y})$, as:

$$\mathbf{y} = \mathbf{k} * \mathbf{x} + \mathbf{n} \text{ with } \mathbf{n} \sim \mathcal{N}(0, \eta^2), \tag{4}$$

where iSTFT denotes inverse short-time Fourier transformation and $*$ is the time-domain linear convolution resulting in $\mathbf{y} \in \mathbb{R}^{L+K-1}$.

Diffusion posterior sampling (DPS) is a technique based on diffusion models that was proposed for solving inverse problems [23] and was recently applied to music restoration tasks [24]. The score function is used as a surrogate speech prior and a log-likelihood term is added to the reverse diffusion, so that the output sample belongs to the posterior $p(\mathbf{X}|\mathbf{Y})$. The unconditional score $\nabla_{\mathbf{X}_\tau} \log p(\mathbf{X}_\tau)$ in (3) is then replaced by the score of the posterior, in order to include the measurement model in the sampling process:

$$\nabla_{\mathbf{X}_\tau} \log p(\mathbf{X}_\tau|\mathbf{Y}) = \nabla_{\mathbf{X}_\tau} \log p(\mathbf{X}_\tau) + \nabla_{\mathbf{X}_\tau} \log p(\mathbf{Y}|\mathbf{X}_\tau). \tag{5}$$

For sampling, a trained score model $s_\theta(\mathbf{X}_\tau, \sigma(\tau)) \approx \nabla_{\mathbf{X}_\tau} \log p(\mathbf{X}_\tau)$ is needed, as well as an approximation of the log-likelihood gradient $\nabla_{\mathbf{X}_\tau} \log p(\mathbf{Y}|\mathbf{X}_\tau)$, since it is generally intractable.

3.2. Log-likelihood approximation

a) Posterior mean approximation:

In [23], the log-likelihood approximation is carried by transferring the outer marginalization with regard to $\mathbf{X}_0$ inside the conditioning, thereby assuming that the *posterior mean* $\mathbb{E}[\mathbf{X}_0|\mathbf{X}_\tau]$ is a sufficient statistic for $\mathbf{X}_\tau$ when modelling the likelihood function:

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}_\tau) &= \int p(\mathbf{Y}|\mathbf{X}_0) p(\mathbf{X}_0|\mathbf{X}_\tau) \mathrm{d}\mathbf{X}_0 \\ &\approx p(\mathbf{Y}| \underbrace{\int \mathbf{X}_0 p(\mathbf{X}_0|\mathbf{X}_\tau) \mathrm{d}\mathbf{X}_0}_{\mathbb{E}[\mathbf{X}_0|\mathbf{X}_\tau]}). \end{aligned} \tag{6}$$

The posterior mean is obtained via the Tweedie formula [25], and can be approximated using our score function estimator :

$$\widehat{\mathbf{M}}(\mathbf{X}_\tau) = \mathbf{X}_\tau + \sigma^2(\tau) \nabla_{\mathbf{X}_\tau} \log p(\mathbf{X}_\tau) \tag{7}$$

$$\approx \mathbf{X}_\tau + \sigma^2(\tau) s_\theta(\mathbf{X}_\tau, \sigma(\tau)). \tag{8}$$

Our measurement model (4) yields the following *posterior mean approximation* for the log-likelihood gradient:

$$\nabla_{\mathbf{X}_\tau} \log p(\mathbf{Y}|\mathbf{X}_\tau) \approx -\frac{1}{\eta^2} \nabla_{\mathbf{X}_\tau} ||\mathbf{y} - \mathbf{k} * \widehat{\mathbf{m}}(\mathbf{X}_\tau)||_2^2, \tag{9}$$

where $\widehat{\mathbf{m}}(\mathbf{X}_\tau) := \mathrm{iSTFT}(\widehat{\mathbf{M}}(\mathbf{X}_\tau))$ and η is the measurement noise level. The resulting reverse probability flow ODE is:

$$\mathrm{d}\mathbf{X}_\tau = -\tfrac{1}{2} g(\tau)^2 s_\theta(\mathbf{X}_\tau, \sigma(\tau)) \mathrm{d}\tau + \zeta(\tau, \eta) \nabla_{\mathbf{X}_\tau} ||\mathbf{y} - \mathbf{k} * \widehat{\mathbf{m}}(\mathbf{X}_\tau)||_2^2 \mathrm{d}\tau, \tag{10}$$

with $\zeta(\tau, \eta) > 0$ a hyperparameter controlling the importance of the measurement error term. According to (5), its theoretical value should be $\zeta(\tau, \eta) = g(\tau)^2/(2\eta^2)$. In [23] however, this hyperparameter is empirically set to $\zeta(\tau, \eta) = \zeta'(\tau)/||\mathbf{y} - \mathbf{k} * \widehat{\mathbf{m}}(\mathbf{X}_\tau)||_2$ so that the measurement error magnitude itself does not influence the

Algorithm 1 Posterior Sampling Scheme

Input: Corrupted $\mathbf{Y}$, RIR $\mathbf{k}$, Reverse step size $\Delta\tau = -\frac{T}{N}$
Output: Clean speech estimate $\widehat{X}$

Sample initial reverse state $\mathbf{X}_T \sim \mathcal{N}(0, \mathbf{I})$
for $n \in \{N, \dots, 1\}$ **do**
  Get diffusion time $\tau = n\frac{T}{N}$
  Phase 1 – Corrector
  Estimate score $s_\theta(\mathbf{X}_\tau, \sigma(\tau))$
  Sample correction noise $\mathbf{W}_c \sim \mathcal{N}(0, \mathbf{I})$
  Correct estimate:
  $\mathbf{X}_\tau \leftarrow \mathbf{X}_\tau + 2r^2\sigma(\tau)^2 s_\theta(\mathbf{X}_\tau, \sigma(\tau)) + 2r\sigma(\tau)\mathbf{W}_c$
  Phase 2 – Predictor
  Estimate score $s_\theta(\mathbf{X}_\tau, \sigma(\tau))$
  Predict next Euler step:
  $\mathbf{X}_\tau \leftarrow \mathbf{X}_\tau - \frac{1}{2} g(\tau)^2 s_\theta(\mathbf{X}_\tau, \sigma(\tau))\Delta\tau$
  Phase 3 – Posterior
  if StateDPS **then** use state approximation (11):
  $\mathbf{x}^{(\mathrm{int})} = \mathrm{iSTFT}(\mathbf{X}_\tau)$
  if DPS **then** use posterior mean approximation (9):
  $\mathbf{x}^{(\mathrm{int})} = \mathrm{iSTFT}\left(\widehat{\mathbf{M}}(\mathbf{X}_\tau)\right)$
  Add log-likelihood gradient:
  $\mathbf{X}_\tau \leftarrow \mathbf{X}_\tau + \zeta(\tau, \eta)\nabla_{\mathbf{X}_\tau} ||\mathbf{y} - \mathbf{k} * \mathbf{x}^{(\mathrm{int})}||_2^2 \Delta\tau$
Output estimate: $\widehat{\mathbf{X}} = \mathbf{X}_0$

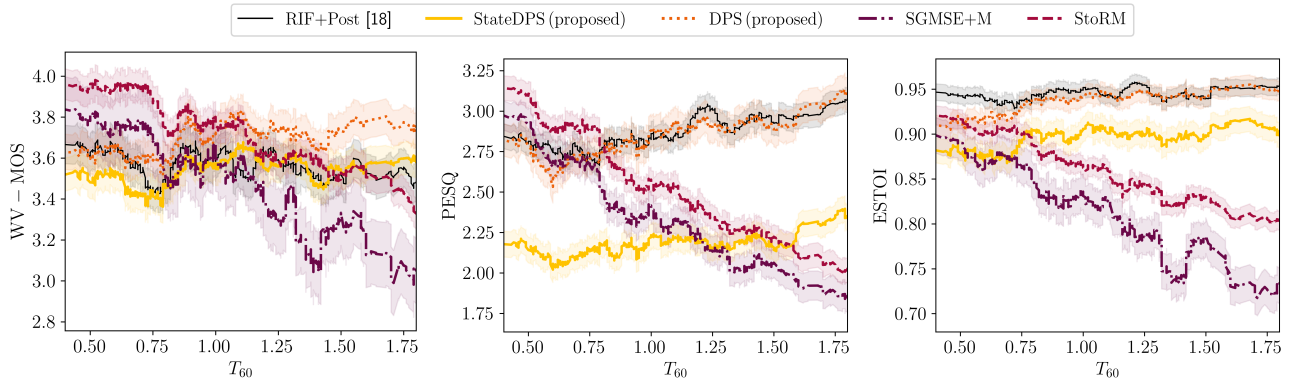


Fig. 1: *Speech metrics as a function of the T_{60} reverberation time. Colored areas cover half a standard deviation range.*

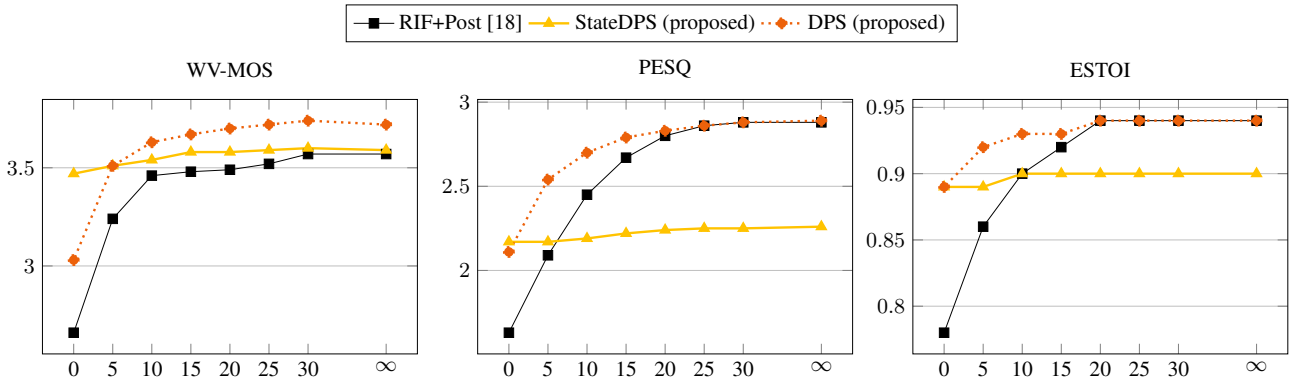


Fig. 2: *Dereverberation performance under zero-mean Gaussian measurement noise. Input SNR indicated on the horizontal axis in* dB.

importance of the gradient step, and $\zeta^{'}(\tau)$ is a schedule which we will describe later in Section 4.2.

b) State approximation:

In [26], a different approximation is used, where the measurement model (4) takes as clean speech reference the current state $\mathbf{X}_\tau$ itself, rather than the posterior mean $\widehat{\mathbf{M}}(\mathbf{X}_\tau)$. This yields the following *state approximation* for the log-likelihood gradient:

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{Y}|\mathbf{X}_\tau) \approx -\frac{1}{\eta^2}\nabla_{\mathbf{x}_\tau}||\mathbf{y} - \mathbf{k} * \mathbf{x}_\tau||_2^2, \qquad (11)$$

with $\mathbf{x}_\tau := \mathrm{iSTFT}(\mathbf{X}_\tau)$. In turn, this results in the following reverse probability flow ODE:

$$\mathrm{d}\mathbf{X}_\tau = -\tfrac{1}{2}g(\tau)^2 s_\theta(\mathbf{X}_\tau, \sigma(\tau))\mathrm{d}\tau + \zeta(\tau,\eta)\nabla_{\mathbf{x}_\tau}||\mathbf{y} - \mathbf{k} * \mathbf{x}_\tau||_2^2\mathrm{d}\tau, \quad (12)$$

this time with $\zeta(\tau,\eta) = \zeta^{'}(\tau)/||\mathbf{y} - \mathbf{k} * \mathbf{x}_\tau||_2$. This approximation becomes less valid as the noise level $\sigma(\tau)$ increases, since for large noise levels, the state $\mathbf{x}_\tau$ is a much worse estimate of $\mathbf{x}_0$ compared to the posterior mean $\widehat{\mathbf{m}}(\mathbf{X}_\tau)$. In practice, the reverse probability flow ODEs (10) and (12) are solved using a predictor-corrector numerical scheme [8] (see Section 4.2).

4. EXPERIMENTAL SETUP

4.1. Data

We generate the WSJ0+Reverb dataset as in [9] in a fashion resembling the WHAMR! dataset recipe [27] by using clean speech data from the WSJ0 dataset and convolving each utterance with a simulated RIR. We use the `pyroomacoustics` package [28] to simulate RIRs. For each utterance, a reverberant room is modeled by sampling uniformly a T_{60} between 0.4 and 1.0 seconds and room dimensions in [5,15]×[5,15]×[2,6] m. This results in an average direct to reverberant ratio (DRR) of -9 dB and average *measured* T_{60} of 0.91 s. An anechoic (but auralized) version of the room is used to generate the reference clean speech, created using the same geometric parameters as the reverberant room but with the absorption coefficient set to 0.99.

4.2. Hyperparameters and training configuration

4.2.1. *Data representation*

When training the unconditional score model, we use only the anechoic part of the generated WSJ0+Reverb data, as the model is supposed to learn the score over clean speech. Utterances are transformed using a short-time Fourier transform (STFT) with a window size of 510 points, a hop length of 128 points and a square-root Hann window, at a sampling rate of 16kHz. In contrast to [9, 19], no compression of the magnitude is used, in order to avoid instabilities when backpropagating small measurement errors through a non-linearity not differentiable around 0. For training, segments of 256 STFT frames (≈2s) are randomly extracted from the utterances and normalized by the maximum absolute value of the segment before feeding them to the network. Using publicly available code, the blind diffusion models SGMSE+ [9] and StoRM [10] are trained on the reverberant and anechoic speech datasets, as these methods are trained in a supervised setting. In comparison, the proposed method does not need any reverberant speech during training.

4.2.2. Forward and reverse diffusion

We set the extremal noise levels for the diffusion schedule g in (1) to $\sigma_{\min} = 0.05$ and $\sigma_{\max} = 0.5$, and the terminal diffusion time to $T = 1$. 50 time steps are used for reverse diffusion with Algorithm 1, which is adapted from the predictor-corrector scheme [8] with probability flow ODE sampling and one step of annealed Langevin dynamics correction with step size of $r = 0.4$.

Tuning $\zeta^{'}(\tau)$ is quite difficult, as setting a high $\zeta^{'}$ leads to pre-echoes, feedback tones and other non-causality artifacts generated by the log-likelihood gradient. Using a low $\zeta^{'}$, however, puts too much emphasis on unconditional generation, therefore increasing the difference between the estimate and the original clean speech. We notice that using the Annealed Langevin Dynamics corrector proposed in [8] helps reduce the aforementioned artifacts and thus provides a more flexible tuning of $\zeta^{'}$. We propose a saw-tooth schedule for $\zeta^{'}(\tau)$ where unconditional speech generation is promoted in the beginning of the reverse process and measurement importance is low towards the end of the process to avoid instabilities:

$$\zeta^{'}(\tau) = \begin{cases} \frac{2500\tau}{0.9} \text{ if } \tau \leq 0.9 \\ \frac{2500(1-\tau)}{0.1} \text{ if } \tau \geq 0.9 \end{cases} \quad (13)$$

4.2.3. Network architecture

The unconditional score network architecture is NCSN++M [10,20], a lighter variant of the NCSN++ [8] which uses ~ 27.8M parameters instead of the original 65M. At each step τ, the current state $\mathbf{X}_\tau$ real and imaginary channels are stacked and fed to the network, and the noise level $\sigma(\tau)$ is provided as a conditioner.

4.2.4. Training configuration

For training the unconditional score model, we use the Adam optimizer with a learning rate of 10^{-4} and an effective batch size of 16 for 300 epochs. We track an exponential moving average of the DNN weights with a decay of 0.999 to be used for sampling as in [9]. A minimal diffusion time is set to $\tau_\epsilon = 0.03$ during training to avoid singularities very close to $\tau = 0$.

4.2.5. Evaluation metrics

For instrumental evaluation of the speech dereverberation performance, we use the intrusive Perceptual Evaluation of Speech Quality (PESQ) [29] and Extended Short-Term Objective Intelligibility (ESTOI) [30] for assessment of speech quality and intelligibility respectively. We also use the non-intrusive WV-MOS [31][2], a DNN-based mean opinion score (MOS) approximation used in [10, 20, 31] for reference-free assessment of bandwidth extension and speech enhancement performance.

5. RESULTS AND DISCUSSION

5.1. Comparison to baselines

In Figure 1, we compare the proposed informed diffusion-based sampling schemes to the informed regularized inverse filtering plus post-processing baseline [18], denoted in the following as RIF+Post. We further add comparisons to the blind dereverberation diffusion methods [9, 10]. Instrumental results are shown as a function of the input T_{60} reverberation time. While the performance of the blind dereverberation methods decreases as T_{60} increases, making the task more difficult, we observe that the informed methods exhibit consistent performance for all considered reverberation times. We notice that the proposed DPS method achieves better or comparable instrumental performance compared to the RIF+Post method [18], while StateDPS performs overall poorer. This shows that using a denoised estimate to match the measurement model, as in the posterior mean approximation (9), increases the dereverberation performance as compared to the state approximation (11). Furthermore, the proposed DPS performed better in terms of subjective quality in informal listening tests: we refer the reader to the audio examples provided in our demo website (see link in abstract). As most diffusion schemes, the proposed (State)DPS methods and the baselines SGMSE+M and StoRM require multiple calls to the score network. Therefore, their computational burden is substantially superior to that of RIF+Post, which is a simple inverse filtering method with real-time capable post-processing.

[2] https://github.com/AndreevP/wvmos

5.2. Robustness to measurement error

In Figure 2, we investigate the robustness of the informed dereverberation approaches to Gaussian measurement noise, added on top of the reverberant speech $\mathbf{y}$. We notice that the proposed DPS is significantly more robust to the introduced noise than the RIF+Post method. This is likely because the prior learned over clean speech by the score model helps gain robustness to mismatches in the measurement model. Informal experiments also show that the degradation of RIF+Post performance to real recorded environmental noise is dramatic, while DPS maintains a very high dereverberation performance and simply lets noise pass through. This shows that the proposed method is much more reliable in realistic scenarios where various noise sources arise, as the remaining noise after DPS dereverberation can easily be removed by a post-processing stage.

5.3. Extension to blind dereverberation

An important aspect of the presented work is that the proposed diffusion posterior sampling technique for informed dereverberation can be extended to blind dereverberation using [32]. In [32] a framework for joint estimation of the blurring kernel and target image is developed using parallel diffusion processes. Our work lays the ground for future adaptation of [32] to jointly estimate the RIR and clean speech in subsequent work. The method we have presented here is interpretable due to the explicit forward model, in contrast to blind dereverberation approaches such as [9, 10]. If successfully implemented, the future method would combine this interpretability with a generative estimation of RIRs. This would furthermore dispose of the need for plug-in RIR estimators in blind scenarios, which the baseline method [18] in contrast requires.

6. CONCLUSIONS AND FUTURE WORK

We have presented a single-channel informed dereverberation method based on diffusion models. The proposed method uses a clean speech prior parameterized by a score model as well as a log-likelihood approximation to generate anechoic speech that fits the measurement model. The approach outperforms an existing state-of-the-art frequency-domain method in terms of robustness to both white Gaussian and real environmental measurement noises. One of the introduced sampling schemes also largely outperforms existing diffusion-based blind dereverberation methods for long reverberation times. The work at hand lays ground to an interpretable extension to blind dereverberation using joint estimation of the RIR and anechoic speech with diffusion models.

7. REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Springer, 2011, vol. 59.

[2] T. Gerkmann and E. Vincent, *Spectral Masking and Filtering*. John Wiley & Sons, 2018.

[3] 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.

[4] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 25, no. 7, pp. 1492–1501, 2017.

[5] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *Proc. Euro. Signal Proc. Conf. (EUSIPCO)*, Sept. 2019.

[6] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 23, no. 6, pp. 982–992, 2015.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Neural Information Proc. Systems (NIPS)*, Dec. 2020.

[8] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Int. Conf. Learning Repr. (ICLR)*, May 2021.

[9] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE Trans. Audio, Speech, Language Proc.*, pp. 1–13, 2023.

[10] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE Trans. Audio, Speech, Language Proc.*, pp. 1–14, 2023.

[11] 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)*, June 2022.

[12] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *The Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 165–169, 07 1979.

[13] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 36, no. 2, pp. 145–152, 1988.

[14] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP J. Adv. Sig. Proc.*, vol. 2007, Dec. 2007.

[15] J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, June 1982.

[16] A. Mertins, T. Mei, and M. Kallinger, "Room impulse response shortening/reshaping with infinity- and p-norm optimization," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 18, no. 2, pp. 249–259, 2010.

[17] H. Schepker, F. Denk, B. Kollmeier, and S. Doclo, "Robust single- and multi-loudspeaker least-squares-based equalization for hearing devices," *EURASIP J. Aud. Speech and Mus. Proc.*, vol. 2022, pp. 1–14, 06 2022.

[18] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2014.

[19] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Interspeech*, Sept. 2022.

[20] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing discriminative versus diffusion generative models for speech restoration tasks," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, June 2023.

[21] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[22] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[23] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," *Int. Conf. Learning Repr. (ICLR)*, May 2023.

[24] E. Moliner, J. Lehtinen, and V. Välimäki, "Solving audio inverse problems with a diffusion model," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, June 2023.

[25] B. Efron, "Tweedie's formula and selection bias," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1602–1614, 2011.

[26] S. Shoushtari, J. Liu, and U. S. Kamilov, "DOLPH: Diffusion models for phase retrieval," *arXiv*, Nov. 2022.

[27] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2020.

[28] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Apr. 2018.

[29] 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.

[30] J. Jensen and C. 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.

[31] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: a unified framework for bandwidth extension and speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, June 2023.

[32] H. Chung, J. Kim, S. Kim, and J. C. Ye, "Parallel diffusion models of operator and image for blind inverse problems," *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

5.2 Blind Dereverberation and Room Impulse Response Estimation with Unsupervised Diffusion Models [P12]

Abstract

In this paper, we present an unsupervised single-channel method for joint blind dereverberation and room impulse response estimation, based on posterior sampling with diffusion models. We parameterize the reverberation operator using a filter with exponential decay for each frequency subband, and iteratively estimate the corresponding parameters as the speech utterance gets refined along the reverse diffusion trajectory. A measurement consistency criterion enforces the fidelity of the generated speech with the reverberant measurement, while an unconditional diffusion model implements a strong prior for clean speech generation. Without any knowledge of the room impulse response nor any coupled reverberant-anechoic data, we can successfully perform dereverberation in various acoustic scenarios. Our method significantly outperforms previous blind unsupervised baselines, and we demonstrate its increased robustness to unseen acoustic conditions in comparison to blind supervised methods. Audio samples and code are available online.

Reference

Eloi Moliner, Jean-Marie Lemercier, Simon Welker, Timo Gerkmann and Vesa Välimäki, "BUDDy: Single-Channel Blind Unsupervised Dereverberation with Diffusion Models", *Int. Workshop on Acoustic Echo and Noise Cancellation (IWAENC)*, Aalborg, Denmark, 2024. DOI: 10.1109/IWAENC61483.2024.10694254

This publication received the Best Student Paper Award at IWAENC 2024

Copyright Notice

Authors' Contributions

The two first authors Eloi Moliner and Jean-Marie Lemercier have equal contribution. They co-wrote Sections 1, 2 and 3. Jean-Marie Lemercier trained all the baselines, conducted several ablations and ran the final speech dereverberation evaluation, he wrote Sections 4 and 5. Eloi Moliner brought significant improvements to the informed baseline in [P10] and developed a large part of the initial BUDDy algorithm. Simon Welker sparked the idea of using the magnitude-compressed STFT L^2-distance, he brought some feedback through discussions on the methods developed in the paper. Timo Gerkmann and Vesa Välimäki brought insights on the experimental validation, mathematical derivations, and reviewed the manuscript.

BUDDY: SINGLE-CHANNEL BLIND UNSUPERVISED DEREVERBERATION WITH DIFFUSION MODELS

Eloi Moliner[1*] *Jean-Marie Lemercier*[2*] *Simon Welker*[2] *Timo Gerkmann*[2] *Vesa Välimäki*[1]

[1] Acoustics Lab, Dept. Information and Communications Eng., Aalto University, Espoo, Finland
[2] Signal Processing (SP), Universität Hamburg, Germany

ABSTRACT

In this paper, we present an unsupervised single-channel method for joint blind dereverberation and room impulse response estimation, based on posterior sampling with diffusion models. We parameterize the reverberation operator using a filter with exponential decay for each frequency subband, and iteratively estimate the corresponding parameters as the speech utterance gets refined along the reverse diffusion trajectory. A measurement consistency criterion enforces the fidelity of the generated speech with the reverberant measurement, while an unconditional diffusion model implements a strong prior for clean speech generation. Without any knowledge of the room impulse response nor any coupled reverberant-anechoic data, we can successfully perform dereverberation in various acoustic scenarios. Our method significantly outperforms previous blind unsupervised baselines, and we demonstrate its increased robustness to unseen acoustic conditions in comparison to blind supervised methods. Audio samples and code are available online[1].

***Index Terms*—** Acoustics, deep learning, speech enhancement

1. INTRODUCTION

When acoustic waves propagate in enclosures and get reflected by walls, the sound received is perceived as reverberated, which can significantly degrade speech intelligibility and quality [1]. The goal of dereverberation is to recover the anechoic component from reverberant speech. We focus here on the single-channel scenario, where measurements from only one microphone are available, which is significantly more challenging than multi-channel scenarios [2].

Traditional dereverberation algorithms assume some statistical properties, such as Gaussianity or sparsity, about the anechoic and reverberant signals. These properties are leveraged to perform dereverberation in the time, spectral or cepstral domain [3]. These methods can tackle *informed* scenarios, where the room impulse response (RIR) is known [4, 5] as well as *blind* scenarios where the RIR is unknown [6, 7]. Informed dereverberation is easier than blind dereverberation, but most scenarios in real-life applications are blind, as the RIR is either not measured beforehand, or becomes invalid even with the slightest deviations in receiver or emitter positions.

Data-driven approaches rely less on such assumptions but rather learn the signal properties and structures from data [8]. Most of these methods are based on supervised learning using pairs of anechoic and reverberant speech. Supervised predictive models have been widely used for blind dereverberation, including time-frequency (T-F) maskers [9], time-domain methods [10] and spectro-temporal mapping [11]. Generative models represent another category of dereverberation algorithms aiming to learn the distribution of anechoic speech conditioned on reverberant input. Some blind supervised methods using generative models such as diffusion models [12, 13] have been recently proposed [14, 15]. However, supervised approaches struggle with limited generalization to diverse acoustic conditions due to the scarcity and variability of available RIR data. Unsupervised approaches offer the potential to circumvent such limitations as they do not require paired anechoic/reverberant data. This paper builds upon prior work [16], which proposed an unsupervised method for informed single-channel dereverberation based on diffusion posterior sampling. The previous study showed the potential of leveraging diffusion models as a strong clean speech prior, which, when combined with a criterion to match the measurement, reached state-of-the-art dereverberation in an informed scenario [16]. This paper extends the method to blind dereverberation, where the unknown RIR is estimated along the anechoic speech. We parameterize the RIR with a model-based subband filter, where each subband of the reverberation filter is modeled by an exponentially decaying signal. The resulting algorithm is an optimization scheme alternating between the diffusion process generating the anechoic speech, and the parameter search estimating the acoustic conditions.

Previous works in related domains explore various parameter estimation techniques for solving blind inverse problems with diffusion posterior sampling. For image deblurring, [17] proposes to use a parallel diffusion process to estimate the deblurring kernel, while [18] adopts an expectation-maximization approach. In the audio domain, [19] addresses the problem of blind bandwidth extension by iteratively refining the parameters of the lowpass filter degradation. Closely related is the work by Saito et al. [20], which performs unsupervised blind dereverberation using DDRM [21] and the weighted-prediction error (WPE) algorithm as initialization [6].

We name our method **BUDD**y for **B**lind **U**nsupervised **D**ereverberation with **D**iffusion Models. We show experimentally that BUDDy efficiently removes reverberation from speech utterances in many acoustic scenarios, thereby largely outperforming previous blind unsupervised techniques. As supervision is not required during the training phase, we demonstrate that BUDDy does not lose performance when presented with unseen acoustic conditions, as opposed to existing blind supervised dereverberation approaches.

2. BACKGROUND

2.1. Diffusion-Based Generative Models

Diffusion-based generative models, or simply *diffusion models* [12, 22], emerged as a class of generative models that learn complex data

[*]These authors contributed equally to this work. The authors gratefully acknowledge the computing resources provided by both the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (NHR project F101AC1) and the Aalto Science-IT project.

[1] uhh.de/sp-inf-buddy.

distributions via iterative denoising. At training time, the target data distribution is transformed into a tractable Gaussian distribution by a *forward process*, incrementally adding noise. During the inference, the *reverse process* refines an initial noise sample into a data sample, by progressively removing noise. The reverse diffusion process, which transports noise samples from a Gaussian prior to the data distribution p_{data}, can be characterized by the following *probability flow* ordinary differential equation (ODE):

$$\mathrm{d}\mathbf{x}_\tau = [\mathbf{f}(\mathbf{x}_\tau, \tau) - \tfrac{1}{2} g(\tau)^2 \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)]\mathrm{d}\tau, \tag{1}$$

where τ indexes the diffusion steps flowing in reverse from $T_{\max}$ to 0. The current diffusion state $\mathbf{x}_\tau$ starts from the initial condition $\mathbf{x}_{T_{\max}} \sim \mathcal{N}(0, \sigma(T_{\max})^2\mathbf{I})$ and ends at $\mathbf{x}_0 \sim p_{\text{data}}$. We adopt the variance exploding parameterization of Karras et al. [23], where the *drift* and *diffusion* are defined as $f(\mathbf{x}_\tau, \tau) = 0$ and $g(\tau) = \sqrt{2\tau}$, respectively. Similarly, we adopt $\sigma(\tau) = \tau$ as the noise variance schedule, which defines the so-called *transition kernel* i.e. the marginal densities: $p_\tau(\mathbf{x}_\tau|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_\tau; \mathbf{x}_0, \sigma(\tau)^2\mathbf{I})$. The *score function* $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)$ is intractable at inference time as we do not have access to $\mathbf{x}_0$. In practice, a *score model* parameterized with a deep neural network $\mathbf{s}_\theta(\mathbf{x}_\tau, \tau)$ is trained to estimate the score function using a *denoising score matching* objective [24].

2.2. Diffusion Posterior Sampling for Dereverberation

Single-channel dereverberation can be considered as the inverse problem of retrieving the anechoic utterance $\mathbf{x}_0 \in \mathbb{R}^L$ from the reverberant measurement $\mathbf{y} \in \mathbb{R}^L$, which is often modelled by convolving the anechoic speech with an RIR $\mathbf{h} \in \mathbb{R}^{L_\mathbf{h}}$, expressed as $\mathbf{y} = \mathbf{h} * \mathbf{x}_0$. We aim to solve this inverse problem by sampling from the posterior distribution $p(\mathbf{x}_0|\mathbf{y}, \mathbf{h})$ of anechoic speech given the measurement and the RIR. We adopt diffusion models for this posterior sampling task by replacing the score function $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)$ in (1) by the *posterior score* $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau|\mathbf{y}, \mathbf{h})$ [13]. Applying Bayes' rule, the posterior score is obtained as

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau|\mathbf{y}, \mathbf{h}) = \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau) + \nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau, \mathbf{h}), \tag{2}$$

where the first term, or *prior score*, can be approximated with a trained score model $\mathbf{s}_\theta(\mathbf{x}_\tau, \tau) \approx \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)$. The likelihood $p(\mathbf{y}|\mathbf{x}_\tau, \mathbf{h})$ is generally intractable because we lack a signal model for $\mathbf{y}$ given the diffusion state $\mathbf{x}_\tau$. We will introduce in the next section a series of approximations to make its computation tractable.

3. METHODS

3.1. Likelihood Score Approximation

In order to obtain a tractable likelihood computation, we posit as in [25] that a one-step denoising estimate of $\mathbf{x}_0$ at time τ can serve as a sufficient statistic for $\mathbf{x}_\tau$ in this context, i.e. that $p(\mathbf{y}|\mathbf{x}_\tau, \mathbf{h}) \approx p(\mathbf{y}|\hat{\mathbf{x}}_0, \mathbf{h})$. Such estimate $\hat{\mathbf{x}}_0$ can be obtained using the score model:

$$\hat{\mathbf{x}}_0 \triangleq \hat{\mathbf{x}}_0(\mathbf{x}_\tau, \tau) = \mathbf{x}_\tau - \sigma(\tau)^2 \mathbf{s}_\theta(\mathbf{x}_\tau, \tau). \tag{3}$$

Furthermore, we consider here that the convolution model remains valid when using this denoised estimate, and therefore that $p(\mathbf{y}|\hat{\mathbf{x}}_0, \mathbf{h}) \approx p(\mathbf{y}|\hat{\mathbf{x}}_0 * \mathbf{h})$. Previous work [16] assumed the error between $\mathbf{y}$ and $\mathbf{h} * \hat{\mathbf{x}}_0$ followed a zero-mean Gaussian distribution, approximating the likelihood score with a simple weighted L^2-distance between $\mathbf{y}$ and $\mathbf{h} * \hat{\mathbf{x}}_0$ in the time-domain. However, in this study, we observed better dereverberation performance using the following distance instead:

$$\mathcal{C}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \| S_{\text{comp}}(\mathbf{y})_{m,k} - S_{\text{comp}}(\hat{\mathbf{y}})_{m,k} \|_2^2, \tag{4}$$

where $S_{\text{comp}}(\mathbf{y}) = |\text{STFT}(\mathbf{y})|^{2/3} \exp\{j \angle \text{STFT}(\mathbf{y})\}$ is the magnitude-compressed spectrogram. This compression accounts for the heavy-tailed nature of speech distributions [26]. With this series of approximations, we obtain the following likelihood score:

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau, \mathbf{h}) \approx -\zeta(\tau) \nabla_{\mathbf{x}_\tau} \mathcal{C}(\mathbf{y}, \mathbf{h} * \hat{\mathbf{x}}_0), \tag{5}$$

where $\zeta(\tau)$ is a diffusion-time dependent hyperparameter that adjusts the weight of the likelihood score during sampling. We use the same gradient-scaled $\zeta(\tau)$ parameterization as in [19, 27].

3.2. Reverberation Operator

The employed reverberation operator relies on a subband filtering approximation [28], which is applied within the Short-Time Fourier Transform (STFT) domain. Let $\mathbf{H} := \text{STFT}(\mathbf{h}) \in \mathbb{C}^{N_\mathbf{h} \times K}$ represent the STFT of an RIR $\mathbf{h}$ with $N_\mathbf{h}$ time frames and K frequency bins. Similarly, let $\mathbf{X} \in \mathbb{C}^{M \times K}$, and $\mathbf{Y} \in \mathbb{C}^{M+N_\mathbf{h}-1 \times K}$, denote the STFTs of anechoic $\mathbf{x}_0$ and reverberant $\mathbf{y}$ speech signals, repectively. The subband convolution operation applies independent convolutions along the time dimension of each frequency band:

$$\mathbf{Y}_{m,k} = \sum_{n=0}^{N_h} \mathbf{H}_{n,k} \mathbf{X}_{m-n,k}. \tag{6}$$

In the blind scenario, we need to estimate $\mathbf{H}$, which is an arduous task without knowledge of the anechoic speech. We constrain the space of possible solutions by designing a structured, differentiable RIR prior whose parameters ψ can be estimated through gradient descent. We denote the complete forward reverberation operator, including forward and inverse STFT, as $\mathcal{A}_\psi(\cdot) : \mathbb{R}^L \to \mathbb{R}^L$.

We denote as $\mathbf{A} \in \mathbb{R}^{N_\mathbf{h} \times K}$ and $\mathbf{\Phi} \in \mathbb{R}^{N_\mathbf{h} \times K}$ the RIR magnitudes and phases of $\mathbf{H}$, respectively. We parameterize the magnitude matrix $\mathbf{A}$ as a multi-band exponential decay model defined in $B < K$ frequency bands. Let $\mathbf{A}' \in \mathbb{R}^{N_\mathbf{h} \times B}$ be the subsampled version of $\mathbf{A}$ in the B selected frequency bands. Each frequency band b is characterized by its weight w_b and exponential decay rate α_b, such that the corresponding subband magnitude filter can be expressed as:

$$\mathbf{A}'_{n,b} = w_b e^{-\alpha_b n}. \tag{7}$$

Once the weights and decay rates parameters are estimated, we reconstruct the magnitudes $\mathbf{A}$ by interpolating the subsampled $\mathbf{A}'$ using $\mathbf{A} = \exp(\text{lerp}(\log(\mathbf{A}')))$, where lerp represents linear interpolation of the frequencies.

Given the lack of structure of RIR phases, we perform independent optimization for each phase factor in $\mathbf{\Phi}$. The resulting set of parameters to optimize is therefore $\psi = \{\mathbf{\Phi}, (w_b, \alpha_b)_{b=1,\dots,B}\}$.

After each optimization step, the estimated time-frequency RIR $\mathbf{H}$ is further processed through a projection step:

$$\overline{\mathbf{H}} = \text{STFT}(\delta \oplus \mathcal{P}_{\text{min}}(\text{iSTFT}(\mathbf{H}))). \tag{8}$$

This operation primarily ensures STFT consistency [29] of $\overline{\mathbf{H}}$. We additionally include a projection $\mathcal{P}_{\text{min}}$ that ensures the time domain RIR has minimum phase lag to guarantee a stable inverse filter, using the Hilbert transform method [30]. Finally, to make the direct-to-reverberation ratio only depend on the late reverberation and to

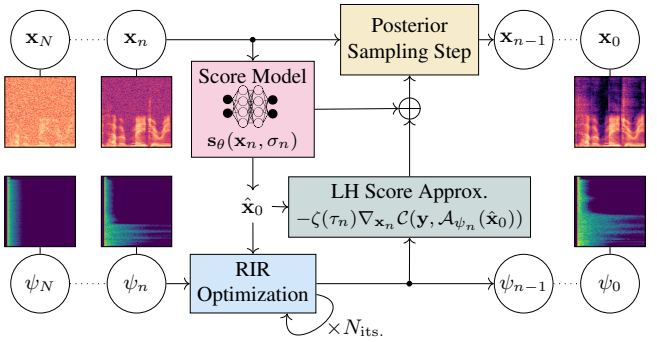


Fig. 1: *Blind unsupervised dereverberation alternating between RIR estimation and posterior sampling for speech reconstruction.*

enforce further constraints on ψ for a more stable optimization, we take the direct path to be at the first sample and with amplitude one. This is achieved by replacing the first sample of the time-domain RIR with a unit impulse, as indicated by the operation $\delta \oplus (\cdot)$.

3.3. Blind Dereverberation Inference

The inference process solves the following objective:

$$\hat{\mathbf{x}}_0, \hat{\psi} = \arg\min_{\mathbf{x}_0, \psi} \mathcal{C}(\mathbf{y}, \mathcal{A}_\psi(\mathbf{x}_0)) + \mathcal{R}(\psi), \quad \text{s.t. } \mathbf{x}_0 \sim p_{\text{data}}. \quad (9)$$

This objective seeks to find the optimal speech $\hat{\mathbf{x}}_0$ and RIR parameters $\hat{\psi}$ that minimize the reconstruction error $\mathcal{C}(\mathbf{y}, \mathcal{A}_\psi(\mathbf{x}_0))$ while also incorporating a regularization term $\mathcal{R}(\psi)$. An essential aspect is the constraint $\mathbf{x}_0 \sim p_{\text{data}}$, which ensures that the estimated signal $\hat{\mathbf{x}}_0$ adheres to the distribution p_{data} of anechoic speech samples. This constraint is implemented in a soft manner by leveraging a pre-trained score model $\mathbf{s}_\theta(\mathbf{x}_\tau, \tau)$ trained on anechoic speech.

The inference algorithm is outlined in Algorithm 1 and visualized in Fig. 1, using the discretization further described in Eq. (11). The algorithm employs the likelihood score approximation from Sec. 3.1, but replacing the convolution with the the reverberation operator $\mathcal{A}_\psi(\cdot)$, while its parameters ψ are optimized in parallel with the speech signal through gradient descent.

We introduce in (9) a *noise regularization* term $\mathcal{R}(\psi)$:

$$\mathcal{R}(\psi) = \frac{1}{N_{\mathbf{h}}} \sum_{l=1}^{N_{\mathbf{h}}} \sum_{k=1}^{K} \| S_{\text{comp}}(\hat{\mathbf{h}}_\psi)_{l,k} - S_{\text{comp}}(\hat{\mathbf{h}}_{\psi'} + \sigma' \mathbf{v})_{l,k} \|_2^2, \quad (10)$$

where $\hat{\mathbf{h}}_\psi = \mathcal{A}_\psi(\delta)$ represents the estimated RIR in the waveform domain, $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a vector of white Gaussian noise, and $\hat{\mathbf{h}}_{\psi'}$ is a copy of the current estimate of $\hat{\mathbf{h}}_\psi$, such that the $\arg\min$ in (9) does not apply to it. In code, this is analogous to detaching the gradients of $\hat{\mathbf{h}}_\psi$ using a stop grad operator. We adopt an annealed schedule for the noise level $\sigma'(\tau)$, resembling the score model schedule $\sigma(\tau)$ but with different hyper-parameters. This regularization term injects noise in the RIR parameter gradients, with decreasing noise power, which enables a wider and smoother exploration while allowing for convergence toward the end of the optimization.

Algorithm 1 Inference algorithm

Require: reverberant speech $\mathbf{y}$
- $\mathbf{x}_{\text{init}} \leftarrow \text{WPE}(\mathbf{y})$
- Sample $\mathbf{x}_N \sim \mathcal{N}(\mathbf{x}_{\text{init}}, \sigma_N^2 \mathbf{I})$ ▷ Warm initialization
- Initialize ψ_N ▷ Initialize the RIR parameters
- **for** $n \leftarrow N, \ldots, 1$ **do** ▷ Discrete step backwards
  - $\mathbf{s}_n \leftarrow s_\theta(\mathbf{x}_n, \tau_n)$ ▷ Evaluate score model
  - $\hat{\mathbf{x}}_0 \leftarrow \mathbf{x}_n - \sigma_n^2 \mathbf{s}_n$ ▷ Get one-step denoising estimate
  - $\hat{\mathbf{x}}_0 \leftarrow \text{Rescale}(\hat{\mathbf{x}}_0)$
  - $\psi_{n-1}^0 \leftarrow \psi_n$ ▷ Use the RIR parameters from last step
  - **for** $j \leftarrow 0, \ldots, N_{\text{its.}}$ **do** ▷ RIR optimization
    - $\mathcal{J}_{\text{RIR}}(\psi_{n-1}^j) \leftarrow \mathcal{C}(\mathbf{y}, \mathcal{A}_{\psi_{n-1}^j}(\hat{\mathbf{x}}_0)) + \mathcal{R}(\psi_{n-1}^j)$
    - $\psi_{n-1}^{j+1} \leftarrow \psi_{n-1}^j - \text{Adam}(\mathcal{J}_{\text{RIR}}(\psi_{n-1}^j))$ ▷ Optim. step
    - $\psi_{n-1}^{j+1} \leftarrow \text{project}(\psi_{n-1}^{j+1})$ ▷ Projection step
  - $\psi_{n-1} \leftarrow \psi_{n-1}^M$
  - $\mathbf{g}_n \leftarrow \zeta(\tau_n) \nabla_{\mathbf{x}_n} \mathcal{C}(\mathbf{y}, \mathcal{A}_{\psi_{n-1}}(\hat{\mathbf{x}}_0))$ ▷ LH score approx.
  - $\mathbf{x}_{n-1} \leftarrow \mathbf{x}_n - \sigma_n(\sigma_{n-1} - \sigma_n)(\mathbf{s}_n + \mathbf{g}_n)$ ▷ Update step
- **return** $\mathbf{x}_0$ ▷ Reconstructed audio signal

4. EXPERIMENTAL SETUP

4.1. Data

We use VCTK [34] as clean speech, selecting 103 speakers for training, 2 for validation and 2 for testing. We curate recorded RIRs from various public datasets (please visit our code repository for details). In total we obtain approximately 10,000 RIRs, and split them between training, validation, and testing using ratios 0.9, 0.05, and 0.05, respectively. The training and validation sets are only used to train the baselines which require coupled reverberant/anechoic data. All data is resampled at 16 kHz.

4.2. Baselines

We compare our method BUDDy to several blind supervised baselines such as NCSN++M [31] and diffusion-based SGMSE+ [14] and StoRM [15]. We also include blind unsupervised approaches leveraging traditional methods such as WPE [6] and Yohena et al. [7], as well as diffusion models Saito et al. [20] and GibbsDDRM [33] with code provided by the authors. For WPE, we take 5 iterations, a filter length of 50 STFT frames (400 ms) and a delay of 2 STFT frames (16 ms).

4.3. Hyperparameters and Training Configuration

Data representation: We train the score model $\mathbf{s}_\theta$ using only the anechoic data from VCTK. For training, 4-s segments are randomly extracted from the utterances. Using publicly available code, the blind supervised models NCSN++M [31], SGMSE+ [14] and StoRM [15] are trained using coupled reverberant/anechoic speech, where the reverberant speech is obtained by convolving the anechoic speech from VCTK with the normalized RIRs.

Reverberation operator: For all methods, STFTs are computed using a Hann window of 32 ms and a hop size of 8 ms. For subband filtering, we further employ 50% zero-padding to avoid aliasing artifacts. Given our sampling rate of $f_s = 16$ kHz, this results in $K = 513$ frequency bins. We set the number of STFT frames of our operator to $N_{\mathbf{h}} = 100$ (800 ms). We subsample the frequency scale in $B = 26$ bands, with a 125-Hz spacing between 0 and 1 kHz, a 250-Hz spacing between 1 and 3 kHz, and a 500-Hz spacing between 3 and 8 kHz. We optimize the RIR parameters ψ with Adam, where the learning rate is set to 0.1, the momentum parameters to $\beta_1 = 0.9$, and $\beta_2 = 0.99$, and $N_{\text{its.}} = 10$ optimization iterations per diffusion step. We constrain the weights w_b between 0 and 40 dB,

Table 1: *Dereverberation results obtained on VCTK-based reverberant datasets. Values indicate mean and standard deviation. We indicate for each method in the table if is blind (i.e. have no knowledge of the RIR) and/or unsupervised. Boldface numbers indicate best performance for supervised and unsupervised methods separately. For all metrics, higher is better.*

			Matched			Mismatched		
Method	Blind	Unsup.	DNS-MOS	PESQ	ESTOI	DNS-MOS	PESQ	ESTOI
Reverberant	-	-	3.14 ± 0.52	1.61 ± 0.37	0.50 ± 0.14	3.05 ± 0.47	1.57 ± 0.29	0.47 ± 0.11
RIF+Post [5]	✗	✓	3.41 ± 0.47	2.66 ± 0.40	0.76 ± 0.09	3.55 ± 0.45	2.86 ± 0.31	0.78 ± 0.09
InfDerevDPS [16]	✗	✓	3.91 ± 0.35	3.77 ± 0.41	0.83 ± 0.09	3.92 ± 0.32	3.69 ± 0.31	0.84 ± 0.08
NCSN++M [31]	✓	✗	3.75 ± 0.38	2.85 ± 0.55	0.80 ± 0.10	3.61 ± 0.39	2.08 ± 0.47	0.64 ± 0.09
SGMSE+M [14, 31]	✓	✗	3.88 ± 0.32	2.99 ± 0.48	0.78 ± 0.09	3.74 ± 0.34	2.48 ± 0.47	$\mathbf{0.69 \pm 0.09}$
StoRM [15]	✓	✗	$\mathbf{3.90 \pm 0.33}$	$\mathbf{3.33 \pm 0.48}$	$\mathbf{0.82 \pm 0.10}$	$\mathbf{3.83 \pm 0.32}$	$\mathbf{2.51 \pm 0.53}$	0.67 ± 0.09
Yohena and Yatabe [7]	✓	✓	2.99 ± 0.56	1.80 ± 0.33	0.55 ± 0.12	2.94 ± 0.44	1.71 ± 0.29	0.51 ± 0.10
WPE [32]	✓	✓	3.24 ± 0.54	1.81 ± 0.42	0.57 ± 0.14	3.10 ± 0.48	1.74 ± 0.37	0.54 ± 0.12
Saito et al. [20]	✓	✓	3.22 ± 0.56	1.68 ± 0.40	0.51 ± 0.13	3.12 ± 0.52	1.70 ± 0.33	0.52 ± 0.10
GibbsDDRM [33]	✓	✓	3.33 ± 0.53	1.70 ± 0.37	0.51 ± 0.13	3.30 ± 0.52	1.75 ± 0.36	0.52 ± 0.11
BUDDy (proposed)	✓	✓	$\mathbf{3.76 \pm 0.41}$	$\mathbf{2.30 \pm 0.53}$	$\mathbf{0.66 \pm 0.12}$	$\mathbf{3.74 \pm 0.38}$	$\mathbf{2.24 \pm 0.54}$	$\mathbf{0.65 \pm 0.12}$

and the decays α_b between 0.5 and 28. This prevents the optimization from approaching degenerate solutions at early sampling stages. Furthermore, we rescale the denoised estimate $\hat{\mathbf{x}}_0$ at each step to match the empirical dataset standard deviation $\sigma_{\text{data}} = 5 \cdot 10^{-2}$, so as to enforce a constraint on the absolute magnitudes of $\hat{\mathbf{h}}_\psi$ and $\hat{\mathbf{x}}_0$.

Forward and reverse diffusion We set the extremal diffusion times to $T_{\max} = 0.5$ and $T_{\min} = 10^{-4}$. For reverse diffusion, we follow Karras et al. [23] and employ a discretization of the diffusion time axis using $N = 200$ steps according to:

$$\forall n < N,\ \tau_n = \sigma_n = \left(T_{\max}^{1/\rho} + \frac{n}{N-1} (T_{\min}^{n/\rho} - T_{\max}^{1/\rho}) \right)^{\rho}, \quad (11)$$

with warping $\rho = 10$. We use the second-order Euler-Heun stochastic sampler in [23] with $S_{\text{churn}} = 50$ and $\zeta' = 0.5$ (prior scaling, see [27]), and the initial point $\mathbf{x}_{\text{init}}$ is taken to be the output of WPE [6] (with same parameters as the WPE baseline) plus Gaussian noise with standard deviation $\sigma = T_{\max}$. The annealing schedule $\sigma'(\tau)$ in the noise regularization term in (10) is the same as the diffusion noise schedule $\sigma(\tau)$ but we bound it between extremal values $\sigma'_{\text{min}} = 5 \times 10^{-4}$ and $\sigma'_{\text{max}} = 10^{-2}$.

Network architecture: To remain consistent with [16], the unconditional score network architecture is NCSN++M [15, 31], a lighter variant of the NCSN++ [13] with 27.8M parameters instead of 65M.

Training configuration: We adopt Adam as the optimizer to train the unconditional score model, with a learning rate of 10^{-4} and an effective batch size of 16 for 190k steps. We track an exponential moving average of the DNN weights with a decay of 0.999.

Evaluation metrics: We assess the quality and intelligibility of speech using the intrusive Perceptual Evaluation of Speech Quality (PESQ) [35] and extended short-term objective intelligibility (ESTOI) [36]. We also employ the non-intrusive DNS-MOS [37], as a DNN-based mean opinion score (MOS) approximation.

5. RESULTS AND DISCUSSION

Table 1 shows the dereverberation results for all baselines and indicates whether each approach is blind and/or unsupervised. We included the results for RIF+Post [5] and InfDerevDPS [16] in the informed scenario to show the upper bound of dereveberation quality one can achieve with perfect knowledge of the room acoustics. We use the same score model $\mathbf{s}_\theta$ and cost function $\mathcal{C}(\cdot, \cdot)$ for InfDerevDPS [16] as for BUDDy. Blind supervised approaches NCSN++M, SGMSE+M, and StoRM largely profit from the supervision during training, and boast a better performance compared to the unsupervised methods. However, in the mismatched setting, their performance dwindles because of their limited generalizability. In contrast, the proposed method BUDDy benefits from unsupervised training, and therefore, modifying the acoustic conditions does not impact performance at all: typically NCSN++M loses 0.78 PESQ by switching from the matched case to the mismatched case, where BUDDy loses 0.06. Our method then outperforms NCSN++M and comes within reach of other supervised approaches, although the generative nature of SGMSE+ and StoRM allow them to retain a relatively high generalization ability. We also observe that the traditional blind unsupervised methods such as WPE [6] and Yohena and Yatabe [7] can only perform limited dereverberation, as they do not benefit from the strong anechoic speech prior that learning-based methods parameterized with deep neural networks offer. Finally, we note that BUDDy performs significantly better on all metrics than the diffusion-based blind unsupervised baselines Saito et al. [20] and GibbsDDRM [33], as these perform mild dereverberation in the presented acoustic conditions, where the input direct-to-reverberant ratio is significanty lower than in the authors' setup.

6. CONCLUSIONS

This paper presents BUDDy, the first unsupervised method simultaneously performing blind dereverberation and RIR estimation using diffusion posterior sampling. BUDDy significantly outperforms traditional and diffusion-based unsupervised blind approaches. Unlike blind supervised methods, which often struggle with generalization to unseen acoustic conditions, our unsupervised approach overcomes this limitation due to its ability to adapt the reverberation operator to a broad range of room impulse responses. While blind supervised methods outperform our approach when the tested conditions match those at training time, our method is on par or even outperforms some supervised baselines in a mismatched setting.

7. REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, vol. 59, Springer, 2011.

[2] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE TASLP*, vol. 36, no. 2, pp. 145–152, 1988.

[3] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

[4] J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in *Proc. ICASSP*, 1982.

[5] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *Proc. ICASSP*, 2014.

[6] T. Nakatani et al., "Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model," *IEEE TASLP*, vol. 16, pp. 1512–1527, 2008.

[7] F. Yohena and K. Yatabe, "Single-channel blind dereverberation based on rank-1 matrix lifting in time-frequency domain," in *Proc. ICASSP*, 2024.

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

[9] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM TASLP*, vol. 25, no. 7, pp. 1492–1501, 2017.

[10] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM TASLP*, vol. 28, pp. 1598–1607, 2020.

[11] K. Han et al., "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM TASLP*, vol. 23, no. 6, pp. 982–992, 2015.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, 2020.

[13] Y. Song et al., "Score-based generative modeling through stochastic differential equations," in *Proc. ICLR*, 2021.

[14] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM TASLP*, vol. 31, pp. 2351–2364, 2023.

[15] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM TASLP*, vol. 31, pp. 2724–2737, 2023.

[16] J.-M. Lemercier, S. Welker, and T. Gerkmann, "Diffusion posterior sampling for informed single-channel dereverberation," in *Proc. WASPAA*, 2023.

[17] H. Chung, J. Kim, S. Kim, and J. C. Ye, "Parallel diffusion models of operator and image for blind inverse problems," *CVPR*, 2023.

[18] C. Laroche, A. Almansa, and E. Coupeté, "Fast diffusion em: a diffusion model for blind inverse problems with application to deconvolution," *IEEE/CVF WACV*, 2024.

[19] E. Moliner, F. Elvander, and V. Välimäki, "Blind audio bandwidth extension: A diffusion-based zero-shot approach," *arXiv*, 2024.

[20] K. Saito et al., "Unsupervised vocal dereverberation with diffusion-based generative models," in *Proc. ICASSP*, 2023.

[21] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Proc. NeurIPS*, 2022.

[22] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. NeurIPS*, 2019.

[23] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Proc. NeurIPS*, 2022.

[24] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[25] H. Chung et al., "Diffusion posterior sampling for general noisy inverse problems," in *Proc. ICLR*, 2023.

[26] T. Gerkmann and R. Martin, "Empirical distributions of dft-domain speech coefficients based on estimated speech variances," *IWAENC*, 2010.

[27] E. Moliner, J. Lehtinen, and V. Välimäki, "Solving audio inverse problems with a diffusion model," in *Proc. ICASSP*, 2023.

[28] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE TASLP*, vol. 15, no. 4, pp. 1305–1319, 2007.

[29] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single channel speech enhancement: History and recent advances," *IEEE Signal Process. Magazine*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

[30] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, 1975.

[31] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing discriminative versus diffusion generative models for speech restoration tasks," in *Proc. ICASSP*, 2023.

[32] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *Proc. ICASSP*, 2008, pp. 85–88.

[33] N. Murata and Sothers, "GibbsDDRM: A partially collapsed Gibbs sampler for solving blind inverse problems with denoising diffusion restoration," in *Proc. ICML*, 2023, pp. 25501–25522.

[34] C. Valentini-Botinhao et al., "Reverberant speech database for training speech dereverberation algorithms and TTS models," *University of Edinburgh*, 2016.

[35] 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 *Proc. ICASSP*, 2001.

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

[37] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*, 2021.

6

Discussion and Conclusions

6.1 Analysis of the Contributions

In this chapter, we summarize and analyze the main contributions of this dissertation. We follow the structure outlined in Section 1.6, starting with predictive model-based algorithms combining signal model approximations and traditional algorithmic structures with DNNs. We then present our main results for speech dereverberation using supervised diffusion models, and conclude with our proposed unsupervised method using both diffusion models and model-based processing.

Model-based Speech Dereverberation

Our first contribution is to highlight the complementarity of the two processing stages proposed in [P1]. We show that the first processing stage based on WPE benefits the post-filtering stage by decorrelating the target speech component and the residual reverberation. Indeed, by integrating the WPE computations in the first stage's DNN optimization graph, we force the DNN to adapt its anechoic speech variance estimation, such that most of the reverberation present in the WPE filter range is removed. This does not necessarily benefit direct speech enhancement metrics such as, e.g. POLQA [P1, comparing E2Ep-WPE and DNN-WPE on Figures 2 and 3]. Informal listening even suggested that fine-tuning the DNN with the proposed end-to-end objective could worsen the perceptual rating. This is due to the residual reverberation now being perceived as an echo, rather than being masked by preceding moderate reverberation typically present if the DNN is not fine-tuned.

However, fine-tuning the first stage helps decorrelating the target speech and residual reverberation components, which allows the Wiener post-filtering stage to perform well. Indeed, the Wiener filter is optimal in the least-square sense only if the target and interference components are statistically uncorrelated [120]. Most dereverberation studies assume the direct speech path and reverberation to be uncorrelated, because multiple reflections with independent phase shifts and arrival times make the coherence of the reverberant sound field vanish [75]. However, this assumption only holds if the mixing time, i.e. the duration between the direct path and the reverberation tail, is sufficiently large. Moreover, in our framework, we define the target speech as the anechoic speech convolved with 40ms and 16ms of the RIR for the hearing-aided and cochlear-implanted scenarios, respectively. Therefore, the target speech and the late reverberation components are likely not to be sufficiently decorrelated for the Wiener filter assumption to hold. Although no formal correlation test is run in this work, we notice empirically that the performance of the Wiener post-filter was highly increased

by fine-tuning the first WPE stage, as reflected by instrumental metrics [P1, comparing E2Ep-WPE+DNN-PF and DNN-WPE+DNN-PF on Figures 2 and 3].

Furthermore, if we remove the first linear filtering stage in the proposed two-stage approach, yielding a simple DNN-assisted single-channel post-filter, severe distortions can be observed in the resulting target speech [P2, Figure 4]. These distortions do not arise with the proposed method, because the WPE algorithm explicitly introduces a prediction delay in order to protect the target speech component [21]. We can draw a parallel to the case of an MVDR beamformer extracting a target speaker in a Gaussian noise field: it can be shown there that the output of the MVDR is a sufficient statistic for further post-filtering [280], i.e. that no information about the target speech is lost when performing MVDR beamforming. We do not dispose of a comparable proof in our case, yet we believe this helps motivate the rationale of using the WPE filter as a non-aggressive means of removing reverberation for further post-filtering.

Another advantage of using a model-based approach over a pure DNN method is that we have an explicit control over several parameters at inference time. That includes the WPE filter length, which we can adjust depending on the available computational budget as well as the severity of the reverberant condition. Additionally, we have control over the amount of early reflections in the output target with both the DNN training objective and the WPE prediction delay. The resulting algorithm can therefore be tailored to either class of hearing device users, namely hearing-aided listeners that benefit from early reflections [86], or cochlear-implanted listeners that do not benefit from early reflections [87]. Furthermore, the computational complexity of our method is lower than that of, e.g. a single-channel GaGNet [P1, Section 5.6]. Though the RLS-based WPE updates are relatively expensive (as large matrix products are involved), we can parallelize most computations and leverage GPU resources because of the frequency subband structure of the algorithm. Yet, a comparison with more recent light-weight multi-channel DNNs as, e.g. [281] , remains to be drawn.

The advantages brought by our second model-based technique [P4] are more straight-forward. There, we compare the speech dereverberation and denoising performances of deep filters [278], [282] to those of DNN -based time-frequency maskers using the same baseline network architecture. We motivate our approach by reminding that the narrow-band approximation (1.5) is valid for the noisy signal model but not for the reverberant signal model, where the CTF model (1.4) should be used instead. Correspondingly, using a deep filter is more in line with the objective of finding a frequency-subband filter that approximately inverts the CTF model, compared to a time-frequency mask. Our contribution in this chapter is therefore to propose a deep filtering extension of time-frequency masks produced by a state-of-the-art denoising baseline [283] parameterized by a single point-wise convolutional layer. The resulting method increases the instrumental speech enhancement metrics by up to 0.34 POLQA points when using 20 frames in the extension. This remarkable performance improvement is obtained with only 1.2% more floating point operations involved and 0.0001% additional trainable parameters compared to the baseline single-frame architecture. In conclusion, this study showcases that a proper analysis of the signal model can lead to a conceptually simple and computationally efficient solution, in comparison to, e.g. blindly increasing the baseline DNN capacity.

Answers to Research Questions

RQ 1 *Are there any benefits of integrating traditional algorithmic structures in DNN-based frameworks, compared to using pure neural networks trained end-to-end?*

We propose in [P1] a DNN-based real-time capable dereverberation algorithm for hearing device users. Separating our algorithm in two distinct processing stages yields a superior interpretability compared to fully end-to-end dereverberation DNNs. In particular, we highlight the complementarity of the two stages, by showing that the first multi-channel filtering stage decorrelates the target speech and residual reverberation, thereby benefiting the second post-filtering stage. Using a DNN-powered WPE algorithm as the first stage permits to adapt the filter length as a function of the reverberant condition, and to match the target specification with respect to early reflections, for hearing-aided or cochlear-implanted listeners. Finally, the resulting approach surpasses a pure single-channel DNN-based approach for only a fraction of the corresponding computational cost.

RQ 2 *How can one integrate the knowledge of the convolutional signal degradation model into the design of DNN-based dereverberation algorithms?*

In [P4], we observe that the narrow-band approximation (1.5) is valid when the degraded speech contains additive noise but not reverberation, where the CTF approximation (1.4) should be used instead. Therefore, we propose an extension strategy for turning DNN-based time-frequency masks into deep filters. As intuitively expected from the connection to subband filtering, we observe that the resulting deep filters perform better on dereverberation than single-frame masks using the same baseline architecture. This increase in performance comes at a negligible fraction of the overall computational cost.

Supervised Conditional Diffusion Models for Speech Dereverberation

In the second chapter of this thesis, we introduce and motivate the use of diffusion-based generative models for speech restoration tasks such as dereverberation. The first contribution of this chapter is an overview of the state of the art in model-based audio restoration using diffusion models. We argue there that domain knowledge can be injected into diffusion models, through distributional assumptions or signal degradation models. For instance, the noise distribution shaping the transition kernel (1.49) can be modified from a Gaussian to a Gamma in order to better fit the estimation error statistics [284]. The signal measurement model, on the other hand, is explicitly leveraged by external conditioning methods using diffusion models for inverse problem solving [P5, Section "Diffusion Models for Inverse Problems"], which is emphasized in the third chapter of this thesis.

In [P7], we run a task-wise comparison between predictive and diffusion-based generative models. In particular, we study the influence of the measurement model, i.e. the type of speech degradation, on the performance gap between predictive and diffusion-based generative models. Our main contribution is the discovery that predictive models tend to underperform compared to diffusion models on non-additive measurement models such as speech reverberation and bandwidth reduction. In reverberant conditions, predictive models have a tendency toward *over-denoising*: they introduce distortions to the target utterance by agressively removing speech content, in particular in low-energy regions and speech onsets/offsets. As for bandwidth

extension, it must be noted that this is a generative task by nature, as it can be seen as inpainting of the high frequencies in the time-frequency domain. Since the predictive model is trained with a complex time-frequency masking objective, it cannot introduce new frequencies and tends to learn an identity mapping. We therefore adapt the predictive model to perform bandwidth extension, by using a learnt filterbank rather than the STFT, and change the masking objective to a mapping one [P7, AE-NCSNN+]. Yet, the resulting predictive model still largely underperforms compared to the diffusion model on the bandwidth extension task. In contrast, the diffusion-based generative model produces high-quality speech in most situations, but can hallucinate in adverse conditions. These hallucinations consist in breathing, gasping or inarticulate vocal content introduced in regions where a speech utterance is wrongly detected [56]. As another minor contribution of this publication, we provide the first diffusion-based method performing bandwidth extension for various bandwidth reduction factors in an agnostic fashion, i.e. without knowledge of the input bandwidth. Our method outperforms all bandwidth-agnostic baselines, and comes close in comparison to a GAN-based approach trained on each input bandwidth specifically [P7, Table 4].

These observations provide a foundation for our follow-up publication [P8]. In this work, we propose to mitigate the tendency of diffusion models to hallucinate, and to increase their restoration performance by making the conditions at the input of the model easier to handle. Analogous to [P1] in Section 3.1, we build a two-stage algorithm where the first stage significantly facilitates the task of the second stage. The first processing stage is a predictive model removing most of the interferences, at the cost of potentially introducing distortions in the target speech. The second stage is a diffusion-based generative model that is trained to remove the residual degradations and restore the destroyed target speech cues, leveraging its generative knowledge of the clean speech distribution. In order to motivate our approach, we take a more precise view on the differences between predictive and generative modelling through the prism of their training objective. Indeed, predictive models trained with a supervised loss learn a mapping from the degraded utterance to a given mode of the posterior distribution of anechoic speech given degraded speech. For instance, if the training loss is a point-wise L^2 distance, it can be shown that the optimal predictive model will map to the mean of the posterior. Relying on this posterior mean ignores the variability of the posterior distribution. Therefore, fine-grained details like low-intensity regions, speech onsets and offsets, which have the highest natural variability in time-frequency speech data, tend to be smoothed out [P8, Figure 2]. This is what we originally identified as over-denoising in our preliminary study [P7]. On the other hand, conditional generative models estimate the whole posterior distribution, and can therefore capture the natural variability of speech data. However, density estimation is a notoriously arduous task compared to mean regression, and failure cases will produce the aforementioned hallucinations. Since the posterior mean is still a valuable cue about the distribution, we use it as an intermediary step to simplify the conditional generation task. We can illustrate the task of speech restoration as a mapping between a degraded sample and a point in the posterior distribution of clean speech given its degraded version [P8, Figure 5]. In this informal view, the predictive estimation provides a way to reduce the distance that the generative model has to cover.

Therefore, our main contribution here is that we propose in [P8] to start the reverse diffusion from the output of a predictive model, and to condition the process by both the initial prediction and the original degraded signal. The empirical results support this intuition, as we observe fewer hallucinations in the output of the proposed method compared to the single diffusion model [P8, Figure 9]. Furthermore, the overall speech quality and interference

removal performance are far superior to either of the isolated processing stages, especially for dereverberation [P8, Table II]. Finally, the proposed two-stage procedure enables to drastically reduce the amount of computations needed for reverse diffusion compared to the sole diffusion model, by reducing the number of neural network evaluations from 100 to 20 for optimal performance [P8, Figure 6]. Our informal intuition is that, in most cases, the trajectory from the posterior mean to the distribution should be less intricate than the path between the degraded signal and the posterior distribution, given that the posterior mean is an informative statistic about the distribution. Therefore, the diffusion model should be able to approximate this smoother trajectory with fewer interpolation steps of the reverse diffusion process.

More broadly speaking, the proposed stochastic regeneration principle transcends speech denoising and dereverberation and can be transposed to a variety of restoration tasks. This is illustrated by the fact that several studies published after our work also investigate in the direction of stochastic regeneration combining predictive and diffusion-based generative models [285]–[294].

Answers to Research Questions

RQ 3 *How do diffusion-based generative models compare to predictive models in terms of speech restoration performance and generalization to unseen conditions? Is there a dependency on the restoration task at hand?*

Our publication [P5] provides an overview of the current state of the art for audio restoration methods relying on diffusion models. The ablation study in [P6] demonstrates that diffusion models generalize better to unseen noise and reverberation conditions than their predictive counterparts. We study in [P7] the influence of the speech degradation model on the performance gap between predictive and diffusion-based generative models. Our experimental results demonstrate that predictive models perform well on denoising tasks. However, diffusion models yield a significantly larger speech quality on non-additive speech degradation models such as reverberation and bandwidth reduction.

RQ 4 *Can one find a suitable combination of predictive and diffusion-based generative models for speech restoration? What are the potential advantages of such hybrid framework compared to pure predictive or generative modelling?*

Our preliminary work in [P7] leads to the introduction in [P8] of the principle of stochastic regeneration, which consists in using a generative model to restore the artifacts created by an initial predictive modelling stage. We motivate the order and structure of this combination based on the observation that predictive models estimate a mode of the posterior distribution of anechoic speech given the reverberant signal, which is an informative cue that generative methods can leverage for modelling the posterior distribution. The proposed approach yields a superior speech quality than pure predictive modelling and improves upon generative modelling in terms of interference removal and hallucinations. Furthermore, our hybrid algorithm largely accelerates inference compared to the original diffusion-based generative model.

Unsupervised Vocal Dereverberation and Room Acoustics Estimation with Diffusion Models

In the last chapter of this thesis, we present a series of work tackling dereverberation with diffusion models in an unsupervised fashion.

Our first contribution is that we propose to solve informed dereverberation as an inverse problem, leveraging a diffusion-based anechoic speech prior [P10] and full knowledge of the RIR. As mentioned in Section 1.3, dereverberation can be framed as a deconvolution task, i.e. an ill-posed inverse problem. Even when the convolution kernel (i.e. the RIR here) is known, the space of solutions consistent with the measurement model is so large that prior knowledge on the search quantity (i.e. anechoic speech here) is required to regularize the solution. If many unstructured priors have been proposed, e.g. in the imaging literature [107]–[110], these offer poor regularization compared to data-driven priors like, e.g. generative models which are far more informative about the actual search quantity. To our knowledge, we propose the first speech prior using an unconditional diffusion model trained on complex time-frequency spectrograms, as DiffWave [212] and WaveGrad [211] operate in the waveform domain.

We show that the resulting approach provides state-of-the-art informed dereverberation, compared to regularized inverse filtering [104]. In particular, the resulting speech quality is very high, which can be attributed to the expressivity of the diffusion-based prior. The method also inherits from the natural robustness of the diffusion model to Gaussian observation noise, given that diffusion models are intrinsically Gaussian denoisers. We also show in audio examples shared online[1] that our method is robust to non-Gaussian measurement noise such as, e.g. street recordings. Interestingly , in this case, it seems that the observation noise is still present in the dereverberated output but that it hardly affects the dereverberation quality. This suggests that combining this approach with a denoising post-processing step could tackle both dereverberaton and denoising. However, in addition to the fact that this method only works in an informed scenario, the method is not robust to fluctuations in the allegedly known RIR. Such fluctuations can be caused by imperfect RIR measurements or mismatches between the geometrical configurations of the recording and the test situation. This acute sensitivity of informed methods has been largely observed in the literature [13]–[15], and we dedicate a whole section in our submitted work [P11] to a quantitative evaluation of this issue in toy and realistic scenarios.

In our following publication [P12], we improve upon the propose informed algorithm and extend it to the blind case, where no information about the room geometry nor the RIR is known. The corresponding contribution is the first algorithm performing joint dereverberation and RIR estimation in a completely unsupervised setting, largely surpassing the existing unsupervised dereverberation state of the art.

We first improve upon our previous work [P10] by changing the diffusion framework for the anechoic speech prior, using findings from Karras et al. [216]. Furthermore, we change the cost function between the reverberant recording and its reconstruction from a L^2 distance on the waveform to a L^2 distance on magnitude-compressed complex spectrograms. The power-law magnitude compression boosts low-energy components such as high frequency regions of speech signals or late reverberation tails, and also account for the heavy-tailedness of speech distributions [153]. The improvement provided by this introduction of prior knowledge is akin to what is mentioned in the first chapter of the thesis. We observe that both these

[1] https://www.inf.uni-hamburg.de/en/inst/ab/sp/publications/waspaa2023-derevdps.html

modifications tremendously increase the speech quality compared to [P10] in the informed scenario, and temper the sensitivity of the method with respect to the hyper-parameter ζ [P10, Equation 10] controlling the trade-off between prior and reconstruction.

The extension of the informed method to the blind case is then realized by designing a parametric RIR approximation inspired from the statistical model in [60], [259]. We assume the reverberant signal model follows the subband approximation (1.4) and represent the subband RIR magnitudes with frequency-dependent exponential decaying envelopes. This model has several advantages. First, it imposes a structure which represents reverberation tails well, constraining the possible solutions to a reduced search space with a physical meaning. Second, using a time-frequency representation with a decimated frequency axis allows to drastically reduce the number of parameters to estimate compared to direct RIR estimation, which simplifies the optimization process. Third, it makes failure cases easier to diagnose, since errors in the estimated RIR can be to some extent determined through a discriminative analysis of the model parameters. To take a practical example, if the estimated RIRs lack high frequencies, one can lower the spectrogram magnitude compression factor in the reconstruction loss [P12, Equations 5 and 6] to compensate for the observed behaviour. Another option could be to lower bound the frequency weights w_b [P12, Equation 7] to force the introduction of high frequencies. Finally, using the inverse problem formulation with an explicit RIR filter enables the introduction of domain knowledge with respect to, e.g. filtering models or room acoustics. For instance, we can manually force the direct path of the RIR to be represented by a unit impulse at the first sample, or constrain the RIR to be a minimum-phase system [P12, Equation 9]. In our case, we observe that such constraints stabilize the RIR parameter optimization. Interestingly, after running our ablation study with respect to these explicit constraints in [P11, Table 2], we realized that enforcing both the minimum-phase property and the fixed direct path simultaneously was not necessary, but that at least one of these constraints was needed to achieve an optimal performance. We interpret this finding by noticing that both these constraints minimize the group delay of the RIR. Indeed, a minimum-phase system has a phase structure that results in the smallest group delay for a given STFT magnitude, while forcing the fixed path to be at the first sample shifts the whole RIR to the left of the time axis.

Our method empirically outmatches prior unsupervised dereverberation state of the art by a large margin. When comparing to supervised methods like, e.g. [P6]–[P8], we show with instrumental metrics and subjective evaluation that the proposed unsupervised approach generalizes to various acoustic conditions while the supervised baselines lose performance when presented with reverberant environments that were unseen during training. This allows our algorithm to perform on par with strong supervised dereverberation methods in such a mismatched scenario, which is to our knowledge the first time this is achieved by an unsupervised dereverberation method.

Research Questions

RQ5 *Can diffusion models provide a good prior for regularizing the inverse problem of single-channel informed dereverberation? What is the resulting robustness with respect to noise in the reverberant recording and errors in the RIR?*

We introduce in [P10] a diffusion-based anechoic speech prior for solving single-channel informed dereverberation as an inverse problem. The resulting model yields state-of-the-art performance among informed single-channel dereverberation techniques, highlighting the role of the diffusion-based prior. We demonstrate the large robustness of our approach with respect to noise in the reverberant recording. The model handles normally distributed noise particularly well, because of the natural Gaussian denoising abilities of diffusion models. However, the proposed method is very sensitive to errors in the RIR, which is a drawback observed in all informed dereverberation methods.

RQ6 *Can one leverage diffusion models and domain knowledge to jointly estimate the room acoustics and anechoic speech from a single-channel reverberant utterance?*

We propose in [P12] the first algorithm performing joint blind dereverberation and RIR estimation in a completely unsupervised setting. The method extends our prior work [P10] to the blind case where the RIR is not available. The RIR approximation uses a parametric design inspired from a time-domain statistical model of late reverberation tails. Following the CTF model, the frequency-subband RIR magnitude is represented by frequency-dependent exponential decaying envelopes. Using domain knowledge reduces the search space for RIR solutions to a subset with high interpretability, and drastically reduces the number of parameters to estimate compared to a direct RIR estimation in the time domain.

6.2 Outlook for Future Research

This thesis open up many questions, and we explore here a few directions for future research.

Improving Model-based Online Dereverberation

The two-stage model-based online dereverberation algorithm proposed in [P1] has a reasonable tradeoff between performance and computational budget. However, recent studies suggest we could largely improve the adaptation speed and processing latency of our method. For instance, the RLS-based WPE could be extended to its Kalman filter version [P3], [151] and the resulting computational complexity reduced to a linear cost with respect to filter length, following Dietzen et al. [150]. It would be interesting to see if the Householder formulation proposed in [137] can also be extended to a Kalman filtering variant, and to study the resulting tradeoff between stability and dereverberation performance.

We considered several single-channel post-filters as candidates for the second processing stage, including [76], [119] and concluded that the Wiener filter (1.18) provided the best performance when applying proper spectral flooring. However, Zhang et al. [295] recently proposed a single-channel speech distortion weighted Wiener filter which allows to explicitly control the tradeoff between residual reverberation and preservation of speech cues. This could be

particularly interesting when the target application is hearing devices. Indeed, when it comes to understanding speech in noise, cochlear-implanted users are more sensitive to residual noise than hearing-aided and normal-hearing listeners, while the latter are more sensitive to speech distortions [85]. Having such an explicit control over this adjustment would increase the degree of customizability of our algorithm [P1] toward various categories of hearing-device users.

Finally, a recent study proposed a new technique for training recurrent neural networks in a parallelizable fashion [102], making training much faster and more efficient. This was so far prohibited, as the dependency of the update gates on previous hidden states required that recurrent network be unrolled during training. The authors propose the minGRU and minLSTM recurrent architectures, removing the aforementioned dependency in the traditional gated recurrent unit (GRU) [100] and LSTM [101] networks. The transformation from LSTM to minLSTM is shown in the following equations:

LSTM [101]

$$
\begin{aligned}
h(n) &= o(n) \odot \tanh(c(n)) \\
o(n) &= \sigma(\text{Linear}([x(n), h(n-1)])) \\
c(n) &= f(n) \odot c(n-1) + i(n) \odot \tilde{c}(n) \\
f(n) &= \sigma(\text{Linear}([x(n), h(n-1)])) \\
i(n) &= \sigma(\text{Linear}([x(n), h(n-1)])) \\
\tilde{c}(n) &= \tanh(\text{Linear}([x(n), h(n-1)]))
\end{aligned}
$$

$\Longrightarrow$

minLSTM [102]

$$
\begin{aligned}
h(n) &= f'(n) \odot h(n-1) + i'(n) \odot \tilde{h}(n) \\
f(n) &= \sigma(\text{Linear}(x(n))) \\
i(n) &= \sigma(\text{Linear}(x(n))) \\
\tilde{h}(n) &= \text{Linear}(x(n)) \\
f'(n) &= \tfrac{f(n)}{f(n)+i(n)} \\
i'(n) &= \tfrac{i(n)}{f(n)+i(n)}
\end{aligned}
$$

There, σ is a sigmoid activation, $x(n)$ is the input of the (min)LSTM cell at time n, $h(n)$ is the hidden state, $c(n)$ is the cell state and $o(n)$, $f(n)$, $\tilde{c}(n)$, $i(n)$ are respectively the signals at the output, forget, control and input gates (see Figure 1.6). This modification allows the authors to employ the parallel scan algorithm for training [296]. In our case, the recurrent structures of both the LSTM architecture and the RLS-based WPE algorithm currently prevent us from employing this learning algorithm, resulting in a significant conceptual and computational overhead during training. It would be certainly valuable to be able to derive a simplification of the RLS (or Kalman) variant of WPE compatible with parallel scan. This boils down to reducing the vector dependencies in the WPE filter, Kalman gain and inverse correlation matrix in Equations (1.40), (1.41) and (1.42) respectively to the following first-order element-wise recurrent form:

$$v(n) = a(n) \odot v(n-1) + b(n)\,, \tag{6.1}$$

where v is the quantity of interest (i.e. Kalman gain, WPE filter, etc.) and a and b are time-dependent scalar coefficients that do not depend on past values of v. This simplified procedure could then be combined with the minLSTM architecture for fast and efficient end-to-end training using parallel scan.

Latent Diffusion Models for Supervised Dereverberation

As mentioned in Sections 1.4.2 and 6.1, one major weakness of diffusion models is their high computational complexity, which raises scepticisim toward their suitability to real-time applications such as hearing device communications. However, motivated by their remarkable generation quality, researchers multiply their efforts to curb the computational burden of

diffusion models. Numerous solutions have already been discussed in Section 4.1, for instance modifying the process time discretization to reduce the number of reverse diffusion steps [232], [235], [237], shrinking the size of score models via classic techniques [98], [99] or using flow matching to rectify the reverse diffusion strategy [167], [227]. Another approach is to try and design spaces where diffusion models can efficiently model the target distribution at a lower expense. Such domains are often hand-crafted through classical transformations, such as STFT for speech [221] or constant-Q transform for music [210], but they can also be learnt by auto-encoding DNNs, leading to so-called *latent diffusion models*. Latent diffusion models are already trending in image generation [297], audio generation [298]–[300], as well as text-conditioned speech generation and editing [301], [302]. However, their application to speech restoration tasks is arguably under-valued at the moment. Compressing full-band (48kHz) speech with DNNs can lead to a drastic reduction in the bitrate used to encode the speech information. This inevitably comes at a slight expense in terms of speech quality, which is leads to the so-called *rate-distortion tradeoff*. The design of such compression schemes called *neural codecs* is quite advanced, and usually relies on CNN-based auto-encoders trained with a mixture of reconstruction and adversarial losses [303]–[305]. The latent spaces learnt by neural codecs perform efficient speech coding, and could therefore be good candidates for computationally efficient diffusion-based dereverberation.

In addition to the potential acceleration provided by using low-dimensional latents, it could be an interesting research direction to examine the influence of the rate-distortion tradeoff on the performance of latent diffusion models. The main parameter influencing the rate-distortion balance is the dimensionality of the latent codes. On the one hand, an auto-encoder with high-dimensional latents yields a large expressivity and conversely a good reconstruction quality. But this means a diffusion model using such latents must learn a lot of fine-grained information, thereby requiring a lot of parameters and computations. On the other hand, low-dimensional latent spaces contain less information and therefore a smaller diffusion model can be used to optimally learn the corresponding distribution. However, the final speech quality is limited by the poorer reconstruction capacities of the codec decoder. Predicting which end of the compromise is the most favorable to speech dereverberation is not trivial and necessitates further research.

Yet, this exploration does not have to be univariate. Indeed, quantization is often used in latent spaces of neural codecs, and this discretization process introduces another control on the rate-distortion tradeoff. In the case of scalar quantization, this is represented by the bit depth attributed to each latent dimension. For vector quantization [306], the corresponding parameter is the number of codewords used to populate the discretized latent space. Such a bivariate optimization over the latent space dimensionality and quantization coarseness is arguably more complex, however this should theoretically result in a better global optimum.

A technical difficulty is that diffusion processes are originally defined on continuous vector spaces, as the score function is not well defined when working in discrete spaces. But modeling discrete data spaces with diffusion models has been at the center of recent interest [307]–[310]. This is particularly motivated by potential applications to text generation, which is currently dominated by auto-regressive language models [311]. Lou et al. [310] formulate probably the most elegant framework for diffusion in discrete spaces, proposing the following *denoising*

score entropy objective, closely related to denoising score matching (see (1.51)):

$$\mathbb{E}_{z_0\sim p(z_0), z_\tau\sim q_\tau(z_\tau|z_0)}\left[\sum_{z'\neq z_\tau}\lambda(z', z_\tau)\left(s_\theta(z_\tau,\tau)(z') - \frac{q_\tau(z'|z_0)}{q_\tau(z_\tau|z_0)}\log s_\theta(z_\tau,\tau)(z')\right)\right], \tag{6.2}$$

where z represents the variable in the latent space and $q_\tau(z_\tau|z_0)$ is the corresponding transition kernel. The discrete-space score model $s_\theta(z_\tau,\tau)(z')$ is trained to estimate the probability ratio $\frac{q_\tau(z'|z_0)}{q_\tau(z_\tau|z_0)}$, analogous to the continuous score function $\nabla_{z_\tau}\log q_\tau(z_\tau|z_0)$. The authors in [307] propose several distributions for the transition kernel $q_\tau(z_\tau|z_0)$, including a discretized Gaussian, a masking process governed by a Bernoulli distribution, a Poisson law and a uniform distribution over the latent space. This opens up numerous research questions: what is the optimal noise distribution for discrete diffusion? is there an ideal dimensionality vs quantization coarseness balance? how to optimally adapt the model size to a given bitrate? etc. This field is still rather new and we believe it has the potential to unveil numerous interesting findings which could make diffusion models for speech dereverberation smaller, faster and more robust.

Efficient and Acoustics-aware Diffusion for Unsupervised Dereverberation

The unsupervised dereverberation and RIR estimation [P12] presented in Section 5.2 is one of our latest works to date, and as such, we have a variety of potential extensions yet to explore.

In our work, we experimented with two posterior sampling schemes, namely DPS [279] and RED-Diff [312]. However, many new samplers for diffusion-based inverse problems have been proposed since. For instance, the CoDPS method in [313] unifies the frameworks in [279], [314], [315] and proposes to use an explicit Gaussian prior on the data. This leads to surprisingly outstanding reconstruction even when the actual prior is far from Gaussian. Remarkably, this formulation avoids the need for backpropagating through the score model, which greatly reduces the computational and memory burden compared to DPS [279]. Unfortunately, DPS and affiliates do not realize true Bayesian sampling but express the posterior as a weighted sum of the diffusion-based log- prior and an approximation of the log- likelihood function. The mixing parameter between prior and likelihood is notoriously hard to tune and its parameterization is not principled theoretically. Instead, Feng et al. [316] propose a variational inference approach where the posterior is not approximated but uses the natural likelihood computation provided by the diffusion framework [55]. The price to pay, however, is solving the complete reverse diffusion procedure at each sampling step, making it computationally impractical. The authors mend this issue in a follow-up work [317] where an evidence lower bound to the diffusion prior is injected in the variational objective. The absence of hyper-parameter compared to DPS leads to more robustness to mismatched priors and makes the method easier to tune to various scenarios [316]. Yet, one of the weaknesses of all the aforementioned samplers resides in the fact that errors committed at early sampling steps cannot be recovered. The sampler proposed in [318] uses a decoupled noise annealing process, which explicitly allows to account for accumulated errors, however at a greater computational cost. Applying and combining these improved posterior sampling techniques could be a promising research axis for increasing the dereverberation quality and sampling speed of our proposed method [P12].

The final and probably most interesting direction for improving our unsupervised dereverberation method [P12] points toward room acoustics modeling. First off, our proposed RIR estimator still lacks a proper model for early reflections, which are hard to access without at

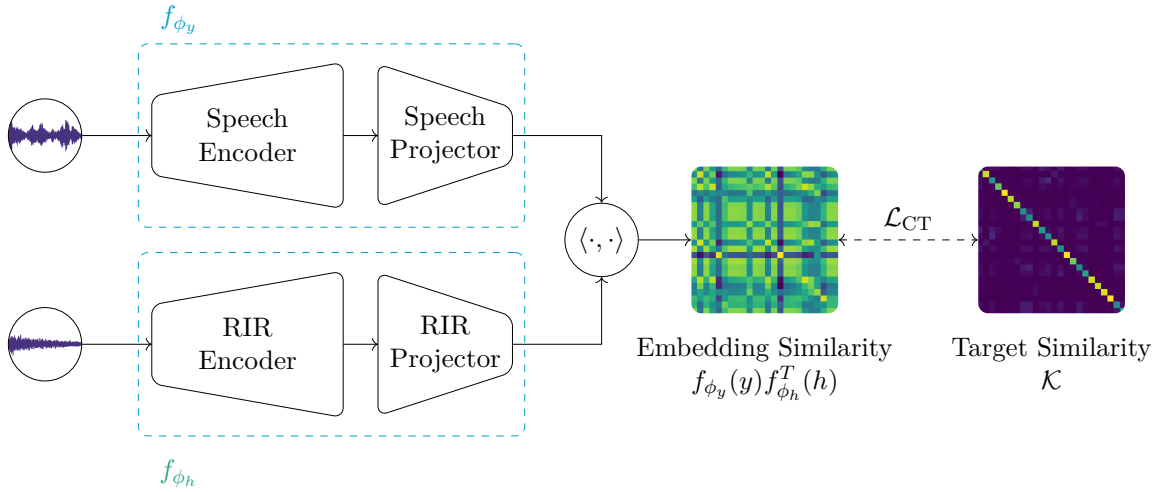


Figure 6.1: *Contrastive learning of shared acoustic embedding. The speech encoder f_{ϕ_y} and RIR encoder f_{ϕ_h} extract embeddings and the resulting embedding similarity matrix is compared to a target similarity matrix $\mathcal{K}$.*

least an approximate knowledge of either the true RIRs [264], [265], [319] or room geometry [263]. Formulating informed methods [263]–[265], [319] in the blind case is a challenging research task. Blind room estimation methods [266], [267], [269]–[272], [274], [277] could prove inspiring in that regard, in particular model-based approaches like the differentiable reverberation approximation proposed in [274].

Another path for better modeling room acoustics lies in self-supervised representations. Seminal works in [258], [320] propose to leverage contrastive learning to obtain high-level embeddings of room acoustics, used for conditioning RIR simulation [257]. The advantage of such representations over aforementioned blind room acoustics estimators is that self-supervised embeddings should benefit from the remarkable generalization abilities of self-supervised learning [321]. Concretely, we propose to modify the RIR regularization term in [P12, eq. 11] to include a loss forcing the embedding of reverberant speech and the estimated RIR to encode the same acoustic representation:

$$\tilde{\mathcal{R}}(\psi) = \mathcal{R}(\psi) + \frac{\langle f_{\phi_y}(y) \cdot f_{\phi_h}(h_\psi) \rangle}{\|f_{\phi_y}(y)\|_2^2 \|f_{\phi_h}(h_\psi)\|_2^2} \tag{6.3}$$

where $\mathcal{R}(\psi)$ is the original noise regularization term in [P12, Equation 11]. The reverberant speech and RIR encoders f_{ϕ_y} and f_{ϕ_h} form the shared acoustic embedding space learnt by contrastive training (see Figure 6.1). The modified objective in (6.3) is straight-forward, but a more intricate question relates to the choice of the contrastive learning objective for training the self-supervised representation. Existing works [257], [258], [320] use a traditional "hard" contrastive objective where positive examples are acoustic scenes corresponding to the exact same room:

$$\text{CrossEntropy}\left(f_{\phi_y}(y) f_{\phi_h}^T(h) \,,\, \mathcal{K} \right), \tag{6.4}$$

with a binary similarity matrix $\mathcal{K}$:

$$\mathcal{K}_{i,j} = \begin{cases} 1 & \text{if room}_i = \text{room}_j \\ 0 & \text{otherwise} \end{cases} \tag{6.5}$$

As mentioned in [322], this strategy might draw apart rooms that are similar but not the same, making the representation over-rely on contrast and not based enough on the characteristics of the room itself. In preliminary work, we have modified the contrastive learning object by parameterizing the similarity matrix between rooms with respect to e.g. subband T_{60} distance:

$$\mathcal{K}_{i,j} = \sum_{f} ||T_{60,\mathrm{f}}(h_i) - T_{60,\mathrm{f}}(h_j)||_2^2 \,. \tag{6.6}$$

Experiments suggested that a T_{60} estimator trained on the resulting representation outperformed the same estimator trained on only reverberant speech spectrograms. However, more time and effort would be needed to draw conclusions on whether this benefits RIR estimation in our proposed method [P12].

References

[1] R. Plomp, "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," *J. Speech Hear. Research*, vol. 29, pp. 146–154, 1986.

[2] H. Dillon, *Hearing Aids*. Boomerang Press, Sydney, Australia, 2001.

[3] V. Hamacher, J. Chalupper, J. Eggers, *et al.*, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP J. Adv. Signal Process.*, vol. 152674, 2005.

[4] D. Wendt, R. Hietkamp, and L. T., "Impact of noise and noise reduction on processing effort: A pupillometry study," *Ear Hear.*, vol. 38, no. 6, pp. 690–700, 2017.

[5] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 331–342, 2006.

[6] B. Champagne, S. Bédard, and A. Stéphenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 4, no. 2, pp. 148–152, 1996.

[7] T. Yoshioka *et al.*, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[8] Q. Jin, Y. Pan, and T. Schultz, "Far-field speaker recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2006.

[9] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-Field Automatic Speech Recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2020.

[10] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Springer, 2011.

[11] S. J. Godsill, P. J. W. Rayner, and O. Cappé, *Digital audio restoration*. Springer, 1998.

[12] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, no. 1, pp. 165–169, 1979.

[13] B. Radlovic, R. Williamson, and R. Kennedy, "Equalization in an acoustic reverberant environment: Robustness results," *IEEE Trans. Audio Speech Lang. Process.*, vol. 8, no. 3, pp. 311–319, 2000.

[14] T. F and W. DB, "Robustness of multichannel equalization in an acoustic reverberant environment," *J. Acoust. Soc. Am.*, vol. 114, no. 2, pp. 833–841, 2003.

[15] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP J. Advances in Signal Process.*, vol. 2007, 2007.

[16] 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.

[17] E. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Eindhoven University of Technology, 2007.

[18] T. Gerkmann, "Cepstral weighting for speech dereverberation without musical noise," in *Proc. Euro. Signal Process. Conf. (EUSIPCO)*, 2011.

[19] B. Yegnanarayana and P. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Audio Speech Lang. Process.*, vol. 8, no. 3, pp. 267–281, 2000.

[20] B. Gillespie, H. Malvar, and D. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2001.

[21] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[22] A. Kuklasiński, S. Doclo, T. Gerkmann, S. Holdt Jensen, and J. Jensen, "Multi-channel PSD estimators for speech dereverberation - a theoretical and experimental comparison," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2015.

[23] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone speech dereverberation, interfering speech cancellation, and noise reduction," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 740–754, 2020.

[24] C. Boeddeker, T. Nakatani, K. Kinoshita, and R. Haeb-Umbach, "Jointly optimal dereverberation and beamforming," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020.

[25] J. Allen, D. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912–915, 1977.

[26] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 1988.

[27] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Audio Speech Lang. Process.*, vol. 11, no. 6, pp. 709–716, 2003.

[28] M. Jeub, M. Schafer, T. Esch, and P. Vary, "Model-Based Dereverberation Preserving Binaural Cues," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1732–1745, 2010.

[29] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 6, pp. 1006–1018, 2015.

[30] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, 2015.

[31] T. Taniguchi, A. S. Subramanian, X. Wang, D. Tran, Y. Fujita, and S. Watanabe, "Generalized Weighted-Prediction-Error Dereverberation with Varying Source Priors For Reverberant Speech Recognition," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2019.

[32] H. Gode, M. Tammen, and S. Doclo, "Joint Multi-Channel Dereverberation and Noise Reduction Using a Unified Convolutional Beamformer With Sparse Priors," in *Proc. ITG Conf. Speech Communication*, 2021.

[33] A. Jukić, N. Mohammadiha, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with low-rank power spectrogram approximation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2015.

[34] K. Sekiguchi, Y. Bando, A. A. Nugraha, M. Fontaine, and K. Yoshii, "Autoregressive Fast Multichannel Nonnegative Matrix Factorization For Joint Blind Source Separation And Dereverberation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021.

[35] H. Kagami, H. Kameoka, and M. Yukawa, "Joint Separation and Dereverberation of Reverberant Mixtures with Determined Multichannel Non-Negative Matrix Factorization," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Calgary, AB, 2018.

[36] D. Schmid, S. Malik, and G. Enzner, "An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2012.

[37] B. Schwartz, S. Gannot, and E. A. P. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 2, pp. 394–406, 2015.

[38] N. Ito, S. Araki, and T. Nakatani, "Probabilistic integration of diffuse noise suppression and dereverberation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2014.

[39] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Speech dereverberation with convolutive transfer function approximation using map and variational deconvolution approaches," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, 2014.

[40] P. Ochieng, "Deep neural network techniques for monaural speech enhancement and separation: State of the art analysis," *Artificial Intelligence Review*, vol. 56, S3651–S3703, 2023.

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

[42] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, 2017.

[43] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 6, pp. 982–992, 2015.

[44] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 1, pp. 102–111, 2017.

[45] L. Bahrman, M. Fontaine, J. L. Roux, and G. Richard, "Speech dereverberation constrained on room impulse response characteristics," in *Proc. Interspeech*, 2024.

[46] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1598–1607, 2020.

[47] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *Proc. Euro. Signal Process. Conf. (EUSIPCO)*, 2019.

[48] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network," in *Proc. Interspeech*, 2018.

[49] W. Ravenscroft, S. Goetze, and T. Hain, "Utterance weighted multi-dilation temporal convolutional networks for monaural speech dereverberation," *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, 2022.

[50] Z.-Q. Wang and D. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 486–490.

[51] Y. Yemini, E. Fetaya, H. Maron, and S. Gannot, "Scene-Agnostic Multi-Microphone Speech Dereverberation," in *Proc. Interspeech*, 2021.

[52] Z.-Q. Wang and D. Wang, "Deep Learning Based Target Cancellation for Speech Dereverberation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 941–950, 2020.

[53] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Neural Inf. Process. Syst.*, 2020.

[54] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. Neural Inf. Process. Syst.*, 2019.

[55] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learning Repr.*, 2021.

[56] 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 Lang. Process.*, vol. 31, pp. 2351–2364, 2023.

[57] K. Prawda, S. J. Schlecht, and V. Välimäki, "Calibrating the sabine and eyring formulas," *J. Acoust. Soc. Am.*, vol. 152, no. 2, pp. 1158–1169, Aug. 2022.

[58] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Int. Conf. Oriental COCOSDA and Conf. Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1–4.

[59] H. Kuttruff, "Room acoustics," *CRC Press*, 2016.

[60] J. Polack, "La transmission de l'énergie sonore dans les salles," Ph.D. dissertation, Université du Maine, Le Mans, 1988.

[61] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, 2007.

[62] E. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, pp. 770–773, 2009.

[63] B. Schwartz, S. Gannot, and E. Habets, "Multi-microphone speech dereverberation using expectation-maximization and Kalman smoothing," in *Proc. Euro. Signal Process. Conf. (EUSIPCO)*, 2013.

[64] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[65] W. C. Sabine, *Collected paper on acoustics*. Harvard University Press, 1921.

[66] J.-M. Jot, "An analysis/synthesis approach to real-time artificial reverberation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 1992.

[67] A. Michael-Titus, P. Revest, and P. Shortland, "Hearing and balance: The auditory and vestibular systems," in *The Nervous System (Second Edition)*, A. Michael-Titus, P. Revest, and P. Shortland, Eds., Churchill Livingstone, 2010, pp. 141–158.

[68] J. van Dorp Schuitman, "Auditory modelling for assessing room acoustics," Ph.D. dissertation, Technische Universiteit Delft, Nov. 2011.

[69] S. Kalafata and K. P. Waye, "The importance of low-frequency masking on auditory perception," Unit for Occupational and Environmental Medicine, Gothenburg University, Tech. Rep., 2020.

[70] W. H. Organization, *World report on hearing*, 2021.

[71] M. Zhang, N. Gomaa, and A. Ho, "Presbycusis: A critical issue in our community," *Int. J. Otolaryngology and Head and Neck Surgery*, vol. 2, no. 4, 2013.

[72] D. Beck, J. Danhauer, H. Abrams, S. Atcherson, D. Brown, M. Chasin, *et al.*, "Audiologic considerations for people with normal hearing sensitivity yet hearing difficulty and/or speech-in-noise problems," *Hearing Review*, vol. 25, no. 10, pp. 28–38, 2018.

[73] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Hayking and K. J. Ray Liu, Eds., Wiley, 2008, pp. 269–302.

[74] S. Doclo, "Multi-microphone noise reduction and dereverberation techniques for speech applications," Ph.D. dissertation, Katholieke Universiteit Leuven, 2003.

[75] E. A. Habets, "Speech dereverberation using statistical reverberation models," in *Speech Dereverberation*, P. Naylor and N. D. Gaubitch, Eds., Springer, 2011, pp. 57–93.

[76] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Audio Speech Lang. Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[77] D. K. Lebart and J.-M. Boucher, "A new method based on spectral subtraction for speech," *Acta Acustica united with Acustica*, no. 3, pp. 359–366, 2001.

[78] M. Dörbecker and S. Ernst, "Combination of two-channel spectral subtraction and adaptive wiener post-filtering for noise reduction and dereverberation," in *Proc. Euro. Signal Process. Conf. (EUSIPCO)*, 1996.

[79] M. Anzalone, L. Calandruccio, K. Doherty, and L. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear Hear.*, vol. 27, no. 5, pp. 480–492, 2006.

[80] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 3029–3038, 2013.

[81] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Am.*, vol. 125, no. 4, pp. 2236–2347, 2009.

[82] K. Kokkinakis, B. Behnam, Y. Hu, and D. R. Friedland, "Single and multiple microphone noise reduction strategies in cochlear implants," *Trends Amplification*, vol. 16, no. 2, pp. 102–116, 2012.

[83] R. Koning, N. Madhu, and J. Wouters, "Ideal time-frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1, pp. 331–341, 2015.

[84] S. J. Mauger, P. W. Dawson, and A. A. Hersbach, "Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 327–336, 2012.

[85] O. ur Rehman Qazi, B. van Dijk, M. Moonen, and J. Wouters, "Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility," *Hearing Research*, vol. 299, pp. 79–87, 2013.

[86] H. S. J. S. Bradley and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am.*, 2003.

[87] Y. Hu and K. Kokkinakis, "Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners," *J. Acoust. Soc. Am.*, 2014.

[88] M. Brændgaard, "Moresound intelligence. oticon tech paper," Oticon, Tech. Rep., 2020.

[89] S. Raufer, P. Kohlhauer, F. Uhlemayr, V. Kühnel, M. Preuss, and S. Hobi, "Spheric speech clarity proven to outperform key competitors for clear speech in noise," Sonova Holding AG, Tech. Rep., 2024.

[90] W. Sumby and I. Pollack, "Visual contributions to speech intelligibility in noise," *J. Acoust. Soc. Am.*, vol. 26, pp. 212–215, 1954.

[91] M. Middelweerd and R. Plomp, "The effect of speechreading on the speech-reception threshold of sentences in noise," *J. Acoust. Soc. Am.*, vol. 82, pp. 2145–2147, 1987.

[92] M. Hay-McCutcheon, D. Pisoni, and K. Hunt, "Audiovisual asynchrony detection and speech perception in hearing-impaired listeners with cochlear implants: A preliminary analysis," *Int. J. Audiol.*, vol. 48, pp. 321–333, 2009.

[93] M. A. Stone and B. C. J. Moore, "Tolerable hearing aid delays. i. estimation of limits imposed by the auditory path alone using simulated hearing losses," *Ear Hear.*, vol. 20, no. 3, pp. 182–192, 1999.

[94] *Clarity challenge: Machine learning challenges for hearing devices*, online.

[95] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *A meeting of the IOC Speech Group on Auditory Modelling at RSRE*, 1987.

[96] D. Mauler and R. Martin, "A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement," in *Proc. Euro. Signal Process. Conf. (EUSIPCO)*, 2007.

[97] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Causal diffusion models for generalized speech enhancement," *IEEE Open Journal of Signal Processing*, 2024.

[98] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," in *Low-Power Computer Vision: Improving the Efficiency of Artificial Intelligence*, G. K. Thiruvathukal, Y.-H. Lu, J. Kim, Y. Chen, and B. Chen, Eds., Chapman & Hall / CRC, 2022.

[99] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NeurIPS Deep Learning Workshop*, 2014.

[100] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Conf. Emp. Methods Nat. Language Proc. (EMNLP)*, 2014.

[101] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[102] L. Feng, F. Tung, M. O. Ahmed, Y. Bengio, and H. Hajimirsadegh, *Were RNNs all we needed?* 2024.

[103] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Neural Inf. Process. Syst.*, 2023.

[104] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2014.

[105] J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 1982.

[106] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Audio Speech Lang. Process.*, vol. 36, no. 2, pp. 145–152, 1988.

[107] T. Chan and C.-K. Wong, "Total variation blind deconvolution," *IEEE Trans. Image Proc.*, vol. 7, no. 3, pp. 370–375, 1998.

[108] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[109] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc.*, vol. 58, no. 1, pp. 267–288, 1996.

[110] B. R. Friend, "Restoring with maximum likelihood and maximum entropy," *J. Opt. Soc. Am.*, vol. 62, pp. 511–518, 1972.

[111] B. Widrow and E. Walach, "Adaptive signal processing for adaptive control," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 1984.

[112] L. Poole, G. Warnaka, and R. Cutter, "The implementation of digital filters using a modified widrow-hoff algorithm for the adaptive cancellation of acoustic noise," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 1984.

[113] A. Mertins, T. Mei, and M. Kallinger, "Room impulse response shortening/reshaping with infinity- and p-norm optimization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 2, pp. 249–259, 2010.

[114] H. Schepker, F. Denk, B. Kollmeier, and S. Doclo, "Robust single- and multi-loudspeaker least-squares-based equalization for hearing devices," *EURASIP J. Advances in Signal Process.*, vol. 2022, pp. 1–14, 2022.

[115] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2006.

[116] J. Mourjopoulos and J. Hammond, "Modelling and enhancement of reverberant speech using an envelope convolution method," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 1983.

[117] T. Langhans and H. Strube, "Speech enhancement by nonlinear multiband envelope filtering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 1982.

[118] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. Sondhi, and Y. Huang, Eds., Springer, 2007, pp. 873–901.

[119] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Letters*, vol. 9, no. 4, pp. 113–119, 2002.

[120] P. Vary and R. Martin, *Digital speech transmission: enhancement, coding and error concealment*. Wiley, 2006.

[121] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1036–1039, 2007.

[122] S. Griebel and M. Brandstein, "Wavelet transform extrema clustering for multi-channel speech dereverberation," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, 1999.

[123] R. Gomez and T. Kawahara, "An improved wavelet-based dereverberation for robust automatic speech recognition," in *Proc. Interspeech*, 2010.

[124] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2002.

[125] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Multi-microphone speech dereverberation using spatio-temporal averaging," in *Proc. Euro. Signal Process. Conf. (EUSIPCO)*, 2004.

[126] B. Cauchi *et al.*, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP J. Advances in Signal Process.*, vol. 2015, pp. 1–12, 2015.

[127] K. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays. Digital Signal Processing*, M. Brandstein and D. Ward, Eds., Springer, 2001, pp. 39–60.

[128] I. Kodrasi and S. Doclo, "Joint Dereverberation and Noise Reduction Based on Acoustic Multi-Channel Equalization," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 4, pp. 680–693, 2016.

[129] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 9, pp. 1879–1890, 2013.

[130] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2008.

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

[132] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori snr estimation approach based on selective cepstro-temporal smoothing," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[133] J. Bitzer and K. Simmer, "Superdirective microphone arrays," in *Microphone Arrays. Digital Signal Processing*, M. Brandstein and D. Ward, Eds., Springer, 2001, pp. 19–39.

[134] N. M. L. A. P. Dempster and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B (Methodological)*, vol. 39, pp. 1–38, 1977.

[135] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, 2012.

[136] F. Yohena and K. Yatabe, "Single-channel blind dereverberation based on rank-1 matrix lifting in time-frequency domain," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024.

[137] J. Wung *et al.*, "Robust multichannel linear prediction for online speech dereverberation using weighted Householder least squares lattice adaptive filter," *IEEE Trans. Signal Process.*, vol. 68, pp. 3559–3574, 2020.

[138] W. Yang, G. Huang, J. Chen, J. Benesty, I. Cohen, and W. Kellermann, "Robust Dereverberation With Kronecker Product Based Multichannel Linear Prediction," *IEEE Signal Process. Letters*, vol. 28, pp. 101–105, 2021.

[139] A. Cohen, G. Stemmer, S. Ingalsuo, and S. Markovich-Golan, "Combined Weighted Prediction Error and Minimum Variance Distortionless Response for dereverberation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017.

[140] T. Nakatani and K. Kinoshita, "A Unified Convolutional Beamformer for Simultaneous Denoising and Dereverberation," *IEEE Signal Process. Letters*, vol. 26, no. 6, pp. 903–907, 2019.

[141] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 1, pp. 69–84, 2011.

[142] M. Fras and K. Kowalczyk, "Convolutional Weighted Parametric Multichannel Wiener Filter for Reverberant Source Separation," *IEEE Signal Process. Letters*, vol. 29, pp. 1928–1932, 2022.

[143] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly Optimal Denoising, Dereverberation, and Source Separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2267–2282, 2020.

[144] T. Nakatani and K. Kinoshita, "Simultaneous Denoising and Dereverberation for Low-Latency Applications Using Frame-by-Frame Online Unified Convolutional Beamformer," in *Proc. Interspeech*, 2019.

[145] J. H. Kim, J. Park, M. Ahn, Y. Lee, W. Kim, and H. Park, "Online Speech Dereverberation Using RLS-WPE Based on a Full Spatial Correlation Matrix Integrated in a Speech Enhancement System," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, 2018.

[146] J. Caroselli, I. Shafran, A. Narayanan, and R. Rose, "Adaptive multichannel dereverberation for automatic speech recognition," in *Proc. Interspeech*, 2017.

[147] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Taipei, Taiwan, 2009.

[148] S. Braun and I. Tashev, "Low Complexity Online Convolutional Beamforming," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2021.

[149] S. Hashemgeloogerdi and S. Braun, "Joint beamforming and reverberation cancellation using a constrained Kalman filter with multichannel linear prediction," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020.

[150] T. Dietzen, S. Doclo, A. Spriet, W. Tirry, M. Moonen, and T. van Waterschoot, "Low-Complexity Kalman filter for multi-channel linear-prediction-based blind speech dereverberation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2017.

[151] S. Braun and E. A. P. Habets, "Online dereverberation for dynamic scenarios using a Kalman filter with an autoregressive model," *IEEE Signal Process. Letters*, vol. 23, no. 12, pp. 1741–1745, 2016.

[152] S. Braun and E. A. P. Habets, "Linear prediction-based online dereverberation and noise reduction using alternating Kalman filters," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 6, pp. 1119–1129, 2018.

[153] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, 2010.

[154] T. Wang, F. Yang, and J. Yang, "Multichannel linear prediction-based speech dereverberation considering sparse and low-rank priors," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 1724–1735, 2024.

[155] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation," in *Proc. Interspeech*, 2017.

[156] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Frame-online DNN-WPE dereverberation," *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, pp. 466–470, 2018.

[157] T. Nakatani *et al.*, "DNN-supported Mask-based Convolutional Beamforming for Simultaneous Denoising, Dereverberation, and Source Separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020.

[158] P. N. Petkov, V. Tsiaras, R. Doddipatla, and Y. Stylianou, "An Unsupervised Learning Approach to Neural-net-supported WPE Dereverberation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019.

[159] J. J. C. Sheeja and B. Sankaragomathi, "Speech dereverberation and source separation using dnn-wpe and lwpr-pca," *Neural Computing and Applications*, vol. 35, pp. 7339–7356, 2023.

[160] Z. Yang, W. Yang, K. Xie, and J. Chen, "Integrating data priors to weighted prediction error for speech dereverberation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 3908–3923, 2024.

[161] S. S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 2002.

[162] M. A. Woodbury, "Inverting modified matrices," Statistical Research Group, Princeton University, Tech. Rep., 1950.

[163] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[164] O. Schwartz, S. Gannot, and E. A. P. Habets, "An Expectation-Maximization Algorithm for Multimicrophone Speech Dereverberation and Noise Reduction With Coherence Matrix Estimation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 9, pp. 1495–1510, 2016.

[165] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[166] S. A. and S. R., "A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends," *Decision Analytics Journal*, vol. 7, p. 100 230, 2023.

[167] A. H. Liu, M. Le, A. Vyas, B. Shi, A. Tjandra, and W.-N. Hsu, "Generative pre-training for speech with flow matching," in *Proc. Int. Conf. Learning Repr.*, 2024.

[168] Y. Zhao, Z.-Q. Wang, and D. Wang, "A two-stage algorithm for noisy and reverberant speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017.

[169] M. Mimura, S. Sakai, and T. Kawahara, "Speech dereverberation using long short-term memory," in *Proc. Interspeech*, 2015.

[170] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Late reverberation suppression using recurrent neural networks with long short-term memory," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2018.

[171] W. Mack, S. Chakrabarty, F.-R. Stöter, S. Braun, B. Edler, and E. Habets, "Single-Channel Dereverberation Using Direct MMSE Optimization and Bidirectional LSTM Networks," in *Proc. Interspeech*, 2018.

[172] A. Purushothaman, D. Dutta, R. Kumar, and S. Ganapathy, "Speech dereverberation with frequency domain autoregressive modeling," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 29–38, 2023.

[173] D. E. Rumelhart, G. E. Hinton, and J. L. McClelland, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, D. E. Rumelhart and J. L. McClelland, Eds., MIT Press, 1987, pp. 318–362.

[174] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[175] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015.

[176] H.-S. Choi, H. Heo, J. H. Lee, and K. Lee, "Phase-aware Single-stage Speech Denoising and Dereverberation with U-Net," in *Proc. Interspeech*, 2020.

[177] L. Zhao, W. Zhu, S. Li, H. Luo, X.-L. Zhang, and S. Rahardja, "Multi-resolution convolutional residual neural networks for monaural speech dereverberation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 2338–2351, 2024.

[178] R. Zhou, W. Zhu, and X. Li, "Speech dereverberation with a reverberation time shortening target," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023.

[179] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, 2012.

[180] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[181] V. Kothapally and J. H. L. Hansen, "Monaural speech dereverberation using deformable convolutional networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 1712–1723, 2024.

[182] H. Wang *et al.*, "Tecanet: Temporal-contextual attention network for environment-aware speech dereverberation," in *Proc. Interspeech*, 2021.

[183] V. Kothapally and J. H. L. Hansen, "Skipconvgan: Monaural speech dereverberation using generative adversarial networks via complex time-frequency masking," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 1600–1613, 2022.

[184] N. Li, M. Ge, L. Wang, and J. Dang, "A fast convolutional self-attention based speech dereverberation method for robust speech recognition," in *Int. Conf. Neural Inf. Process. (ICONIP)*, 2019.

[185] C. Chen, W. Sun, D. Harwath, and K. Grauman, "Learning audio-visual dereverberation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023.

[186] K. Tan, Y. Xu, S.-X. Zhang, M. Yu, and D. Yu, "Audio-Visual Speech Separation and Dereverberation With a Two-Stage Multimodal Network," *IEEE Journ. Selected Topics Signal Process.*, vol. 14, no. 3, pp. 542–553, 2020.

[187] S. Chowdhury, S. Ghosh, S. Dasgupta, A. Ratnarajah, U. Tyagi, and D. Manocha, "Adverb: Visually guided audio dereverberation," in *Int. Conf. Computer Vision (ICCV)*, 2023.

[188] W. Zhang *et al.*, "End-to-End Dereverberation, Beamforming, and Speech Recognition with Improved Numerical Stability and Advanced Frontend," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021.

[189] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified Architecture for Multichannel End-to-End Speech Recognition With Neural Beamforming," *IEEE Journ. Selected Topics Signal Process.*, vol. 11, no. 8, pp. 1274–1288, 2017.

[190] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Joint optimization of neural network-based WPE dereverberation and acoustic model for robust online ASR," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019.

[191] L. Li, Y. Kang, Y. Shi, L. Kürzinger, T. Watzel, and G. Rigoll, "Adversarial joint training with self-attention mechanism for robust end-to-end speech recognition," *EURASIP J. Advances in Signal Process.*, vol. 2021, 2021.

[192] M. Togami, "End To End Learning For Convolutive Multi-Channel Wiener Filtering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021.

[193] Z.-Q. Wang, G. Wichern, and J. L. Roux, "Convolutive prediction for monaural speech dereverberation and noisy-reverberant speaker separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3476–3490, 2021.

[194] G. Carbajal, R. Serizel, E. Vincent, and E. Humbert, "Joint NN-Supported Multichannel Reduction of Acoustic Echo, Reverberation and Noise," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2158–2173, 2020.

[195] B. T. Feng, R. Baptista, and K. L. Bouman, "Neural approximate mirror maps for constrained diffusion models," *arXiv*, 2024.

[196] H. Attias, J. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Proc. Neural Inf. Process. Syst.*, 2000.

[197] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *Proc. Int. Conf. Learning Repr.*, 2014.

[198] D. Baby and H. Bourlard, "Speech dereverberation using variational autoencoders," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021.

[199] P. Wang and X. Li, "RVAE-EM: Generative speech dereverberation based on recurrent variational auto-encoder and convolutive transfer function," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024.

[200] G. C. G. 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.

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

[202] I. Goodfellow *et al.*, "Generative adversarial networks," in *Proc. Neural Inf. Process. Syst.*, vol. 27, 2014.

[203] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv*, 2014.

[204] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," *Int. Conf. Computer Vision (ICCV)*, pp. 2813–2821, 2016.

[205] K. Wang, J. Zhang, S. Sun, Y. Wang, F. Xiang, and L. Xie, "Investigating generative adversarial networks based speech dereverberation for robust speech recognition," in *Proc. Interspeech*, 2018.

[206] J. Su, Z. Jin, and A. Finkelstein, "Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *Proc. Interspeech*, 2020.

[207] H. Muckenhirn *et al.*, "Cyclegan-based unpaired speech dereverberation," in *Proc. Interspeech*, 2022.

[208] D. P. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," in *Proc. Neural Inf. Process. Syst.*, 2021.

[209] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Neural Inf. Process. Syst.*, vol. 34, 2021.

[210] E. Moliner, J. Lehtinen, and V. Välimäki, "Solving audio inverse problems with a diffusion model," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023.

[211] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," *Proc. Int. Conf. Learning Repr.*, 2021.

[212] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," *Proc. Int. Conf. Learning Repr.*, 2021.

[213] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications.* Journal of the American Statistical Association, 2000, vol. 82.

[214] B. D. Anderson, "Reverse-time diffusion equation models," *Stoch. Proc. and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[215] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[216] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Proc. Neural Inf. Process. Syst.*, 2022.

[217] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. Cambridge University Press, 2007.

[218] E. Moliner and V. Välimäki, "Diffusion-based audio inpainting," *J. Audio Eng. Soc.*, vol. 72, pp. 100–113, 3 Mar. 2024.

[219] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *Proc. Asia-Pacific Signal and Information Processing Association (APSIPA)*, 2021.

[220] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022.

[221] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Proc. Interspeech*, 2022.

[222] R. Kimura *et al.*, "Diffusion model-based mimo speech denoising and dereverberation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. Workshops (ICASSPW)*, 2024.

[223] T. Nakatani, N. Kamo, M. Delcroix, and S. Araki, "Multi-stream diffusion model for probabilistic integration of model-based and data-driven speech enhancement," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, 2024.

[224] T. Nakatani, B. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 8, pp. 1512–1527, 2008.

[225] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *Proc. Neural Inf. Process. Syst.*, 2018.

[226] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. Int. Conf. Machine Learning*, 2015.

[227] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *Proc. Int. Conf. Learning Repr.*, 2023.

[228] A. A. Nugraha, K. Sekiguchi, and K. Yoshii, "A flow-based deep latent variable model for speech spectrogram modeling and enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1104–1117, 2020.

[229] M. Strauss and B. Edler, "A flow-based neural network for time domain speech enhancement," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 5754–5758, 2021.

[230] C. Villani, *Optimal Transport: old and new*. Springer, 2009, vol. 338.

[231] E. Moliner, S. Braun, and H. Gamper, "Gaussian Flow Bridges for Audio Domain Transfer with Unpaired Data," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, 2024.

[232] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learning Repr.*, 2022.

[233] Z. Kong and W. Ping, "On fast sampling of diffusion probabilistic models," in *Proc. Int. Conf. Machine Learning*, 2021.

[234] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *Proc. Int. Conf. Learning Repr.*, 2022.

[235] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, and Y. Ren, "Prodiff: Progressive fast diffusion model for high-quality text-to-speech," in *ACM Multimedia*, 2022.

[236] R. San-Roman, E. Nachmani, and L. Wolf, "Noise estimation for generative diffusion models," *arXiv*, 2021.

[237] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "BDDM: Bilateral denoising diffusion models for fast and high-quality speech synthesis," in *Proc. Int. Conf. Learning Repr.*, 2022.

[238] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," in *Proc. Int. Conf. Machine Learning*, 2023.

[239] G. Daras, Y. Dagan, A. G. Dimakis, and C. Daskalakis, "Consistent diffusion models: Mitigating sampling drift by learning to be consistent," in *Proc. Neural Inf. Process. Syst.*, 2023.

[240] B. Lay, J.-M. Lemercier, J. Richter, and T. Gerkmann, "Single and few-step diffusion for generative speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024.

[241] S. Mosayyebpour and F. Nesta, "Neural-Network Supervised Maximum Likelihood-based on-line Dereverberation," in *Proc. Euro. Signal Process. Conf. (EUSIPCO)*, 2018.

[242] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (RED)," *SIAM J. Imag. Sci.*, vol. 10, no. 4, pp. 1804–1844, 2017.

[243] L. Drude *et al.*, "Integrating neural network based beamforming and weighted prediction error dereverberation," in *INTERSPEECH*, 2018.

[244] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017.

[245] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, p. 100 258, 2022.

[246] P. Srivastava, A. Deleforge, and E. Vincent, "Realistic sources, receivers and walls improve the generalisability of virtually-supervised blind acoustic parameter estimators," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022.

[247] online.

[248] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[249] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2018.

[250] S. Sakamoto, A. Ushiyama, and H. Nagatomo, "Numerical analysis of sound propagation in rooms using the finite difference time domain method," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 3003–3008, 1989.

[251] A. Krokstad, S. Strom, and S. Sorsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *J. Sound and Vibration*, vol. 8, no. 1, pp. 118–125, 1968.

[252] M. Vorländer, "Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm," *J. Acoust. Soc. Am.*, vol. 86, no. 1, pp. 172–178, 1989.

[253] Z. Tang, L. Chen, B. Wu, D. Yu, and D. Manocha, "Improving reverberant speech training using diffuse acoustic simulation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020.

[254] W. Wittebol, H. Wang, M. Hornikx, and P. Calamia, "A hybrid room acoustic modeling approach combining image source, acoustic diffusion equation, and time-domain discontinuous galerkin methods," *Applied Acoustics*, vol. 223, p. 110 068, 2024.

[255] P. Masztalski, M. Matuszewski, K. Piaskowski, and M. Romaniuk, "StoRIR: Stochastic room impulse response generation for audio data augmentation," in *Proc. Interspeech*, 2020.

[256] A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha, and D. Yu, "Fast-rir: Fast neural diffuse room impulse response generator," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022.

[257] F. Lluìs and N. Meyer-Kahlen, "Blind spatial impulse response generation from separate room- and scene-specific information," *arXiv*, 2024.

[258] P. Götz, C. Tuna, A. Walther, and E. A. P. Habets, "Contrastive representation learning for acoustic parameter estimation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023.

[259] J. A. Moorer, "About this reverberation business," *Computer music journal*, pp. 13–28, 1979.

[260] R. Ratnam, D. L. Jones, B. C. Wheeler, J. O'Brien William D., C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, Oct. 2003.

[261] H. W. Löllmann and P. Vary, "Estimation of the reverberation time in noisy environments," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, 2008.

[262] H. Löllmann, A. Brendel, and W. Kellermann, "Comparative Study of Single-Channel Algorithms for Blind Reverberation Time Estimation," in *Int. Cong. Acoustics*, (Aachen), Sep. 9–13, 2019.

[263] S. Dilungana, A. Deleforge, C. Foy, and S. Faisan, "Geometry-informed estimation of surface absorption profiles from room impulse responses," in *Proc. Euro. Signal Process. Conf. (EUSIPCO)*, 2022.

[264] T. Sprunck, A. Deleforge, Y. Privat, and C. Foy, "Gridless 3d recovery of image sources from room impulse responses," *IEEE Signal Process. Letters*, vol. 29, pp. 2427–2431, 2022.

[265] T. Sprunck, A. Deleforge, Y. Privat, and C. Foy, "Fully reversing the shoebox image source method: From impulse responses to room parameters," 2024.

[266] H. Gamper and I. J. Tashev, "Blind reverberation time estimation using a convolutional neural network," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, 2018.

[267] P. P. Parada, D. Sharma, and P. A. Naylor, "Non-intrusive estimation of the level of reverberation in speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024.

[268] R. Falcon Perez, G. Götz, and V. Pulkki, "Machine-learning-based estimation of reverberation time using room geometry for room effect rendering," in *Int. Cong. Acoustics*, 2019.

[269] S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "Blind estimation of room acoustic parameters and speech transmission index using mtf-based cnns," in *Proc. Euro. Signal Process. Conf. (EUSIPCO)*, 2021.

[270] C. Foy, A. Deleforge, and D. Di Carlo, "Mean absorption estimation from room impulse responses using virtually supervised learning," *J. Acoust. Soc. Am.*, vol. 150, no. 2, pp. 1286–1299, Aug. 2021.

[271] P. Götz, C. Tuna, A. Walther, and E. A. P. Habets, "Online reverberation time and clarity estimation in dynamic acoustic conditions," *J. Acoust. Soc. Am.*, vol. 153, no. 6, pp. 3532–3542, Jun. 2023.

[272] C. J. Steinmetz, V. K. Ithapu, and P. Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2021.

[273] K. Lee, J. Seo, K. Choi, S. Lee, and B. S. Chon, "Room impulse response estimation in a multiple source environment," in *AES Int. Conf. Spat. Immersive Audio*, 2023.

[274] S. Lee, H.-S. Choi, and K. Lee, "Differentiable artificial reverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2541–2556, 2022.

[275] V. Välimäki, B. Holm-Rasmussen, B. Alary, and H.-M. Lehtonen, "Late reverberation synthesis using filtered velvet noise," *Applied Sciences*, vol. 7, no. 5, 2017.

[276] J. P. Stautner and M. Puckette, "Designing multi-channel reverberators," *Computer Music Journal*, vol. 6, p. 52, 1982.

[277] S. Lee, H.-S. Choi, and K. Lee, "Yet another generative model for room impulse response estimation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2023.

[278] W. Mack and E. A. P. Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Process. Letters*, vol. 27, pp. 61–65, 2020.

[279] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," in *Proc. Int. Conf. Learning Repr.*, 2023.

[280] R. Balan and J. P. Rosca, "Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase," in *Sensor Array and Multichannel Signal Proc. Workshop*, 2002.

[281] K. Tesch and T. Gerkmann, "Insights into deep non-linear filters for improved multi-channel speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 563–575, 2023.

[282] H. Schröter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, "DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022.

[283] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, 2022.

[284] E. Nachmani, R. S. Roman, and L. Wolf, "Denoising diffusion gamma models," in *Proc. Int. Conf. Learning Repr.*, 2022.

[285] X. Cao and S. Zhao, "Pfgm++ combined with stochastic regeneration for speech enhancement," in *Int. Conf. Signal and Image Proc. (ICSIP)*, 2024, pp. 267–271.

[286] H. Shi *et al.*, "Diffusion-based speech enhancement with joint generative and predictive decoders," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 12 951–12 955.

[287] N. Kamo, M. Delcroix, and T. Nakatani, "Target speech extraction with conditional diffusion model," in *Proc. Interspeech*, 2023.

[288] R. Sawata *et al.*, "Diffiner: A versatile diffusion-based generative refiner for speech enhancement," in *Proc. Interspeech*, 2023.

[289] H. Wang, J. Villalba, L. Moro-Velazquez, J. Hai, T. Thebaud, and N. Dehak, "Noise-robust speech separation with fast generative correction," in *Proc. Interspeech*, 2024.

[290] S. Wang, S. Liu, A. Harper, P. Kendrick, M. Salzmann, and M. Cernak, "Diffusion-based speech enhancement with schrödinger bridge and symmetric noise schedule," *arXiv*, 2024.

[291] T. Trachu, C. Piansaddhayanon, and E. Chuangsuwanich, "Thunder : Unified regression-diffusion speech enhancement with a single reverse step using brownian bridge," *arXiv*, 2024.

[292] C. Wang, J. Gu, D. Yao, J. Li, and Y. Yan, "Gald-se: Guided anisotropic lightweight diffusion for efficient speech enhancement," *arXiv*, 2024.

[293] Y. Liu *et al.*, "Fadi-aec: Fast score based diffusion model guided by far-end signal for acoustic echo cancellation," *arXiv*, 2024.

[294] D. Kim *et al.*, "Guided conditioning with predictive network on score-based diffusion model for speech enhancement,"

[295] J. Zhang, R. Tao, J. Du, and L.-R. Dai, "Sdw-swf: Speech distortion weighted single-channel wiener filter for noise reduction," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3176–3189, 2023.

[296] F. A. Heinsen, "Efficient parallelization of a ubiquitous sequential computation," *arXiv*, 2023.

[297] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2022.

[298] H. Liu *et al.*, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proc. Int. Conf. Machine Learning*, 2023.

[299] H. Liu, T. Wang, J. Cao, R. He, and J. Tao, *Boosting fast and high-quality speech synthesis with linear diffusion*, 2023.

[300] F. Schneider, "Archisound: Audio generation with diffusion," *ETH Zürich*, 2023.

[301] K. Shen *et al.*, "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *Proc. Int. Conf. Learning Repr.*, 2024.

[302] Z. Ju *et al.*, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," in *Proc. Int. Conf. Machine Learning*, 2024.

[303] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *arXiv*, 2021.

[304] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," in *Proc. Neural Inf. Process. Syst.*, 2022.

[305] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," in *Proc. Neural Inf. Process. Syst.*, 2023.

[306] R. M. Gray, "Vector quantization," *IEEE ASSP Magazine*, 1984.

[307] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. van den Berg, "Structured denoising diffusion models in discrete state-spaces," in *Proc. Int. Conf. Machine Learning*, 2021.

[308] S. Gu *et al.*, "Vector quantized diffusion model for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2022.

[309] A. Campbell, J. Benton, V. D. Bortoli, T. Rainforth, G. Deligiannidis, and A. Doucet, "A continuous time framework for discrete denoising models," in *Proc. Neural Inf. Process. Syst.*, 2022.

[310] A. Lou, C. Meng, and S. Ermon, "Discrete diffusion modeling by estimating the ratios of the data distribution," in *Proc. Int. Conf. Machine Learning*, 2024.

[311] T. B. Brown *et al.*, "Language models are few-shot learners," in *Proc. Neural Inf. Process. Syst.*, 2020.

[312] M. Mardani, J. Song, J. Kautz, and A. Vahdat, "A variational perspective on solving inverse problems with diffusion models," in *Proc. Int. Conf. Learning Repr.*, 2024.

[313] N. Yismaw, U. S. Kamilov, and M. S. Asif, "Gaussian is all you need: A unified framework for solving inverse problems via diffusion posterior sampling," 2024.

[314] H. Chung, S. Lee, and J. C. Ye, "Decomposed diffusion sampler for accelerating large-scale inverse problems," in *Proc. Int. Conf. Learning Repr.*, 2024.

[315] J. Song, A. Vahdat, M. Mardani, and J. Kautz, "Pseudoinverse-guided diffusion models for inverse problems," in *Proc. Int. Conf. Learning Repr.*, 2022.

[316] B. T. Feng, J. Smith, M. Rubinstein, H. Chang, K. L. Bouman, and W. T. Freeman, "Score-based diffusion models as principled priors for inverse imaging," in *Int. Conf. Computer Vision (ICCV)*, 2023.

[317] B. T. Feng and K. L. Bouman, "Variational bayesian imaging with an efficient surrogate score-based prior," *Trans. Machine Learning Res.*, 2024.

[318] B. Zhang, W. Chu, J. Berner, C. Meng, A. Anandkumar, and Y. Song, "Improving diffusion inverse problem solving with decoupled noise annealing," 2024.

[319] T. Rajapaksha, X. Qiu, E. Cheng, and I. Burnett, "Geometrical room geometry estimation from room impulse responses," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2016.

[320] J. Bitterman, D. Levi, H. H. Diamandi, S. Gannot, and T. Rosenwein, "Revrir: Joint reverberant speech and room impulse response embedding using contrastive learning with application to room shape classification," in *Proc. Interspeech*, 2024.

[321] W. Huang, M. Yi, X. Zhao, and Z. Jiang, "Towards the generalization of contrastive self-supervised learning," in *Proc. Int. Conf. Learning Repr.*, 2023.

[322] P. Götz, C. Tuna, A. Brendel, A. Walther, and E. A. P. Habets, "Blind acoustic parameter estimation through task-agnostic embeddings using latent approximations," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, 2024.

List of Acronyms

AR	augmented reality
ASR	automatic speech recognition
ATF	acoustic transfer function
CNN	convolutional neural network
CTF	convolutive transfer function
DNN	deep neural network
DPS	diffusion posterior sampling
DRR	direct-to-reverberant ratio
EDC	energy decay curve
EDR	energy decay relief
EM	expectation-maximization
GAN	generative adversarial network
GMM	Gaussian mixture model
GSC	generalized sidelobe canceller
ISM	image source method
LSTM	long short-term memory
LTI	linear time-invariant
MAP	maximum a posteriori
MCEM	Monte Carlo Expectation Maximization
MFCC	mel-frequency cepstral coefficients
ML	maximum likelihood
MVDR	minimum variance distortionless response
NMF	non-negative matrix factorization
PDF	probability density function
PSD	power spectral density
RIR	room impulse response
RLS	recursive least squares
RNN	recurrent neural network
RTF	real-time factor

SDE	stochastic differential equation
SNR	signal-to-noise ratio
STFT	short-time Fourier transform
TDOA	time difference of arrival
VAE	variational auto-encoder
VR	virtual reality
WPE	weighted prediction error
XAI	explainable artificial intelligence

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

Hamburg, 04.11.2024

Jean-Marie Lemercier

Appendices

A

Related Peer-Reviewed Publications

A.1 Customizable End-To-End Optimization Of Online Neural Network-Supported Dereverberation For Hearing Devices [P2]

Abstract

This work focuses on online dereverberation for hearing devices using the weighted prediction error (WPE) algorithm. WPE filtering requires an estimate of the target speech power spectral density (PSD). Recently deep neural networks (DNNs) have been used for this task. However, these approaches optimize the PSD estimate which only indirectly affects the WPE output, thus potentially resulting in limited dereverberation. In this paper, we propose an end-to-end approach specialized for online processing, that directly optimizes the dereverberated output signal. In addition, we propose to adapt it to the needs of different types of hearing-device users by modifying the optimization target as well as the WPE algorithm characteristics used in training. We show that the proposed end-to-end approach outperforms the traditional and conventional DNN-supported WPEs on a noise-free version of the WHAMR! dataset.

Reference

Jean-Marie Lemercier, Joachim Thiemann, Raphael Koning and Timo Gerkmann "Customizable End-To-End Optimization Of Online Neural Network-Supported Dereverberation For Hearing Devices", *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Singapore, Singapore, 2022, DOI: 10.1109/ICASSP43922.2022.9746235

Copyright Notice

Authors' Contributions

Jean-Marie Lemercier is the first author of this publication. He implemented all algorithms, trained the neural networks used in the paper, conducted the experimental validation, and wrote the manuscript. Timo Gerkmann brought insights on the experimental validation, mathematical derivations, and also thoroughly reviewed the paper. Joachim Thiemann and Raphael Koning brought their feedback on all methods through discussions, they also helped with reviewing the manuscript.

CUSTOMIZABLE END-TO-END OPTIMIZATION OF ONLINE NEURAL NETWORK-SUPPORTED DEREVERBERATION FOR HEARING DEVICES

Jean-Marie Lemercier[⋆]*, Joachim Thiemann*[†]*, Raphael Koning*[†]*, Timo Gerkmann*[⋆]

[⋆]Signal Processing (SP), Universität Hamburg, Germany
[†]Advanced Bionics, Hanover, Germany
{firstname.lastname}@uni-hamburg.de, {firstname.lastname}@advancedbionics.com

ABSTRACT

This work focuses on online dereverberation for hearing devices using the weighted prediction error (WPE) algorithm. WPE filtering requires an estimate of the target speech power spectral density (PSD). Recently deep neural networks (DNNs) have been used for this task. However, these approaches optimize the PSD estimate which only indirectly affects the WPE output, thus potentially resulting in limited dereverberation. In this paper, we propose an end-to-end approach specialized for online processing, that directly optimizes the dereverberated output signal. In addition, we propose to adapt it to the needs of different types of hearing-device users by modifying the optimization target as well as the WPE algorithm characteristics used in training. We show that the proposed end-to-end approach outperforms the traditional and conventional DNN-supported WPEs on a noise-free version of the WHAMR! dataset.

Index Terms— online algorithm, dereverberation, neural network, end-to-end learning, hearing devices

1. INTRODUCTION

Communication and hearing devices require modules aiming at suppressing undesired parts of the signal to improve the speech quality and intelligibility. Reverberation is one of such distortions caused by room acoustics, and is characterized by multiple reflections on the room enclosures. Late reflections particularly degrade the speech signal and may result in a reduced intelligibility [1, 2].

Many traditional approaches were proposed for dereverberation such as spectral enhancement [3], beamforming [4], a combination of both [5], coherence weighting [6, 7, 8], and linear-prediction based approaches such as the weighted-prediction error (WPE) algorithm [9, 10]. WPE computes an auto-regressive multi-channel filter and applies it to a delayed group of reverberant speech frames. The approach is able to cancel late reverberation while preserving early reflections, thus improving speech intelligibility for normal and hearing-aided listeners [11, 12]. WPE and its extensions have been shown to be robust and efficient multi-channel techniques. However, these methods require the prior estimation of the anechoic speech power spectrum density (PSD), which is modelled for instance through the speech periodogram [9], by an autoregressive process [13] or through non-negative matrix factorization [14]. A deep neural network (DNN) was first proposed in [15] to model the anechoic PSD, thus avoiding the use of an iterative refinement process.

This work has been funded by the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380. The authors are responsible for the content of this paper.

As hearing devices require to operate in real-time in variable environments, the methods implemented should be suited for frame-to-frame online processing, as well as being adaptive to changing room acoustics. Online adaptive approaches are based on either Kalman filtering [16, 17] or on a recursive least squares (RLS) adapted WPE. In this latter RLS-WPE framework, the PSD is either estimated by recursive smoothing of the reverberant signal [18] or by a DNN [19].

In the previously cited work, the neural network was trained toward PSD estimation, although the aim of the algorithm is WPE-based dereverberation. End-to-end techniques were proposed, using an Automatic Seech Recognition (ASR) criterion in order to refine the front-end DNN handling e.g. speech separation [20], denoising [21], or multiple tasks [22]. An end-to-end procedure for online dereverberation and ASR based on DNN-WPE was proposed in [23]. However, for hearing devices, it is less clear which criterion reaches optimal speech intelligibility and quality, and such performance is highly dependent on the considered user category.

In this work, we propose to use a criterion on the WPE output short-time spectrum for online dereverberation to improve instrumentally predicted speech intelligibility and quality. To solve the issue of the initialization period of RLS-WPE, we design a dedicated training procedure taking into account the adaptive nature of the algorithm. Finally we include a specialization toward different hearing-device users categories: hearing-aid (HA) users on the one hand benefiting from early reflections like normal listeners [11]; cochlear-implanted (CI) on the other hand which do not benefit from early reflections [24].

The rest of this paper is organized as follows. In Section 2, the online DNN-WPE dereverberation scheme is summarized, followed by a description of the proposed end-to-end training procedure in Section 3. The experimental setup is described in Section 4 and the evaluation results are presented and discussed in Section 5.

2. SIGNAL MODEL AND DNN-SUPPORTED WPE DEREVERBERATION

2.1. Signal model

In the short-time Fourier transform (STFT) domain using the subband-filtering approximation [9], the reverberant speech $\boldsymbol{x} \in \mathbb{C}^D$ is obtained at the D-microphone array by convolution of the anechoic speech s and the room impulse responses (RIRs) $\boldsymbol{H} \in \mathbb{C}^{D\times D}$ with length L,

$$\boldsymbol{x}_{t,f} = \sum_{\tau=0}^{L} \boldsymbol{H}_{\tau,f} s_{t-\tau,f} = \boldsymbol{d}_{t,f} + \boldsymbol{e}_{t,f} + \boldsymbol{r}_{t,f}, \qquad (1)$$

where t denotes the time frame index and f the frequency bin, which

we will drop when not needed. $\boldsymbol{d}$ denotes the direct path, $\boldsymbol{e}$ the early reflections component, and $\boldsymbol{r}$ the late reverberation. The early reflections component $\boldsymbol{e}$ was shown to contribute to speech quality and intelligibility for normal and HA listeners [12] but not for CI listeners, particularly in highly-reverberant scenarios [24]. Therefore, we propose that the dereverberation objective is to retrieve $\boldsymbol{\nu} = \boldsymbol{d} + \boldsymbol{e}$ for HA listeners and $\boldsymbol{\nu} = \boldsymbol{d}$ for CI listeners.

2.2. WPE dereverberation

In relation to the subband reverberant model in (1), the WPE algorithm [9] uses an auto-regressive model to approximate the late reverberation $\boldsymbol{r}$. Based on a zero-mean time-varying Gaussian model on the STFT anechoic speech s with time-frequency dependent PSD $\lambda_{t,f}$, a multi-channel filter $\boldsymbol{G} \in \mathbb{C}^{DK \times D}$ with K taps is estimated. This filter aims at representing the inverse of the late tail of the RIRs $\boldsymbol{H}$, such that the target $\boldsymbol{\nu}$ can be obtained through linear prediction, with a delay Δ avoiding undesired short-time speech cancellations, which also leads to preserving parts of the early reflections:

$$\widehat{\boldsymbol{\nu}}_{t,f} = \boldsymbol{x}_{t,f} - \boldsymbol{G}_{t,f}^H \boldsymbol{X}_{t-\Delta,f}, \tag{2}$$

where $\boldsymbol{X}_{t-\Delta,f} = \left[\boldsymbol{x}_{t-\Delta,f}^T, \ldots, \boldsymbol{x}_{t-\Delta-K+1,f}^T\right]^T \in \mathbb{C}^{DK}$.

In order to obtain an adaptive and real-time capable approach, RLS-WPE was proposed in [18], where the WPE filter $\boldsymbol{G}$ is recursively updated along time:

$$\boldsymbol{K}_{t,f} = \frac{(1-\alpha)\boldsymbol{R}_{t-1,f}^{-1}\boldsymbol{X}_{t-\Delta,f}}{\alpha\lambda_{t,f} + (1-\alpha)\boldsymbol{X}_{t-\Delta,f}^H \boldsymbol{R}_{t-1,f}^{-1}\boldsymbol{X}_{t-\Delta,f}}, \tag{3}$$

$$\boldsymbol{R}_{t,f}^{-1} = \frac{1}{\alpha}\boldsymbol{R}_{t-1,f}^{-1} - \frac{1}{\alpha}\boldsymbol{K}_{t,f}\boldsymbol{X}_{t-\Delta,f}^T \boldsymbol{R}_{t-1,f}^{-1}, \tag{4}$$

$$\boldsymbol{G}_{t,f} = \boldsymbol{G}_{t-1,f} + \boldsymbol{K}_{t,f}(\boldsymbol{x}_{t,f} - \boldsymbol{G}_{t-1,f}^H \boldsymbol{X}_{t-\Delta,f})^H. \tag{5}$$

$\boldsymbol{K} \in \mathbb{C}^{DK}$ is the Kalman gain, $\boldsymbol{R} \in \mathbb{C}^{DK \times DK}$ the covariance of the delayed reverberant signal buffer $\boldsymbol{X}_{t-\Delta,f}$ weighted by the PSD λ, and α the forgetting factor.

2.3. DNN-based PSD estimation

The anechoic speech PSD $\lambda_{t,f}$ is estimated at each time step t, either by recursive smoothing of the reverberant periodogram [18] or with help of a DNN [19]. A block diagram of the DNN-WPE algorithm as proposed in [19] is given in Figure 1. In this approach, the channel-averaged magnitude frame $|\bar{\boldsymbol{x}}_t|$ is fed as input to a recurrent neural network with state h_t and the output is a target speech mask $\mathcal{M}_{t,f}^{(\nu)}$. The PSD estimate is then obtained by time-frequency masking:

$$\widehat{\lambda}_{t,f} = (\mathcal{M}_{t,f}^{(\nu)} \odot |\bar{\boldsymbol{x}}_{t,f}|)^2. \tag{6}$$

The DNN is optimized with a mean-squared error criterion on the masked output in [15, 19]. In contrast, we propose to use the Kullback-Leibler (KL) divergence as it led to better results:

$$\mathcal{L}_{\text{DNN-WPE}} = \text{KL}(\mathcal{M}_{t,f}^{(\nu)} \odot |\bar{\boldsymbol{x}}_{t,f}|, |\boldsymbol{\nu}_{t,f}|). \tag{7}$$

The training objective $\mathcal{L}_{\text{DNN-WPE}}$ does not match the output $\widehat{\boldsymbol{\nu}}$ of the whole algorithm, thus potentially limiting the dereverberation performance.

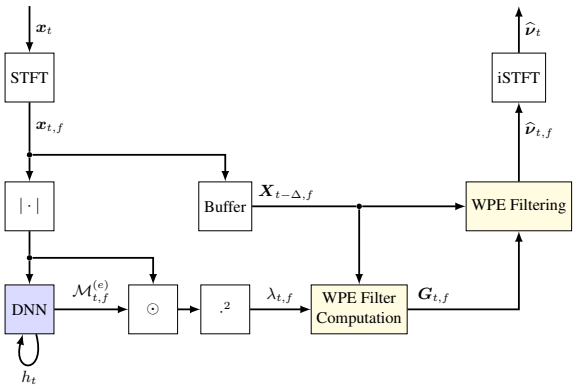


Fig. 1. DNN-supported online WPE dereverberation. Blue blocks refer to trainable neural network layers. Yellow blocks represents adaptive statistical signal processing

3. PROPOSED END-TO-END TRAINING PROCEDURE FOR ONLINE DEREVERBERATION OPTIMALITY

3.1. End-to-end criterion and objectives

Here we propose an end-to-end training procedure where the optimization criterion is placed at the output of the DNN-WPE algorithm. The objective is to include the back-end WPE into the computations through which the loss will be backpropagated during training:

$$\mathcal{L}_{\text{E2E}} = \text{KL}(|\widehat{\boldsymbol{\nu}}_{t,f}|, |\boldsymbol{\nu}_{t,f}|). \tag{8}$$

In contrast to [23], no ASR criterion is used here. Instead, the loss is computed in the time-frequency domain. This enables us to take different targets and WPE parameters into consideration, for customizing the approach towards different hearing-device user categories. Namely, for HA listeners, where early reflections are considered beneficial [12], we set the training target to $\boldsymbol{\nu} = \boldsymbol{d} + \boldsymbol{e}$ and we use a larger prediction delay Δ_{HA}. For CI listeners, for which early reflections may be harmful [24], we set $\boldsymbol{\nu} = \boldsymbol{d}$ and we use a shorter delay $\Delta_{\text{CI}} < \Delta_{\text{HA}}$ to remove as much of the early component as possible given the delayed linear prediction model (5).

3.2. Initialization period

As all operations in RLS-WPE are differentiable, we can use backpropagation through the whole WPE algorithm. However, an important practical aspect of this study focuses on handling the initialization period of the RLS-WPE algorithm. During this interval of L time frames, the filter $\boldsymbol{G}$ has not yet converged to a stable value, and the resulting dereverberation performance is suboptimal, as we will show it in the experiments (see Section 5).

Therefore, rather than relying on a hypothetical shortening of this period through implicit PSD optimization [23], we choose to exclude this initialization period from training, which leads us to design the procedure as given in Algorithm 1. Finally, we investigate using a pretrained DNN, trained on the same dataset with the loss function (7), and plugging it into Algorithm 1 for fine-tuning.

Algorithm 1 End-to-End Training Procedure

Extract STFT of given sequence
Segment sequence in N segments of size L
for $n \in \{0 \dots N-1\}$ **do**

---

  if $n = 0$ **then** ▷ Initialization period
    Initialize LSTM state $h_0^{(0)} = 0$
    Initialize WPE statistics

$$\boldsymbol{G}_{0,f}^{(0)} = \mathbf{0}\,,\ (\boldsymbol{R}^{-1})_{0,f}^{(0)} = \boldsymbol{I}$$

    for $t \in \{0 \dots L-1\}$ **do**
      Compute $\widehat{\boldsymbol{e}}_{t,f}$ with one pass of DNN-WPE

---

  if $n > 0$ **then** ▷ After initialization
    Initialize LSTM state $h_0^{(n)} = h_{L-1}^{(n-1)}$
    Initialize WPE statistics

$$\boldsymbol{G}_{0,f}^{(n)} = \boldsymbol{G}_{L-1,f}^{(n-1)}\,,\ (\boldsymbol{R}^{-1})_{0,f}^{(n)} = (\boldsymbol{R}^{-1})_{L-1,f}^{(n-1)}$$

    for $t \in \{0 \dots L-1\}$ **do**
      Compute $\widehat{\boldsymbol{e}}_{t,f}$ with one pass of DNN-WPE
    Backpropagate loss (8) through time on n
    Repeat [13:] to re-update $h_{L-1}^{(n)}, \boldsymbol{G}_{L-1,f}^{(n)}$

---

4. EXPERIMENTAL SETUP

4.1. Dataset generation

The data generation is inspired from the WHAMR! dataset [25] and uses anechoic speech utterances from the WSJ0 dataset. As the initialization time L typically corresponds to 4 seconds when using a forgetting factor of $\alpha = 0.99$, we concatenate utterances belonging to the same speaker and construct sequences of approximately 20 seconds. Within each sequence, permutations of the utterances are used to create several versions of the sequence, so as not to lose too much data since the first segment is never used for optimization.

These sequences are convolved with 2-channel RIRs generated with the RAZR engine [26] and randomly picked. Each RIR is generated by uniformly sampling room acoustics parameters as in [25] and a T_{60} reverberation time between 0.4 and 1.0 seconds. As target data for the HA case, the first 40 ms of the RIR is convolved with the utterance, representing the direct path and the early reflections, whereas for the CI scenario, only the direct path is retained. Each training set consists of approximately 55 hours of speech data sampled at 16 kHz.

4.2. Hyperparameter settings

All approaches are trained by backpropagating the KL divergence through time, using the Adam optimizer with a learning rate of 10^{-4}, exponentially decreasing by a factor of 0.96 at every epoch. Early stopping with a patience of 10 epochs and mini-batches size of 128 segments are used. The STFT uses a square-rooted Hann window of 32 ms and a 75 % overlap, and segments of $L = 4$ s are constructed.

The WPE filter length is set to $K = 10$ STFT frames (~ 80ms) as our goal is to focus on the beginning of the reverberation tail, where most of the reverberant energy lies. Another reason is that the WPE computational complexity globally increases with the square of K, making end-to-end training longer and more unstable.

	Initialization (4.0 s)				After initialization			
	ELR	PESQ	ESTOI	SDR	ELR	PESQ	ESTOI	SDR
Unprocessed	*2.9*	*2.29*	*0.61*	*4.0*	*2.9*	*2.26*	*0.61*	*3.9*
Oracle-WPE-HA	3.0	2.49	0.65	6.5	**7.6**	**2.83**	**0.77**	**7.0**

Table 1. Oracle WPE dereverberation performance during and after the initialization period. HA scenario. For all metrics, the higher the better. $T_{60} \in [0.4, 1.0]$

The number of channels is $D = 2$, the adaptation factor $\alpha = 0.99$ and the delays $\Delta_{\text{HA}} = 5$ frames for the HA scenario and $\Delta_{\text{CI}} = 3$ frames for the CI scenario. Those delay values are picked as they experimentally provide optimal evaluation metrics when comparing the corresponding target to the output of WPE when using the oracle PSD. This setting allows to obtain a real-time factor - defined as the ratio between the time needed to process an utterance and the length of the utterance - below 0.1 with all computations performed on a Nvidia GeForce RTX 2080Ti GPU. A simple decision criterion is used to prevent WPE from updating filter values when the input speech power goes below -30 dB, corresponding to speech pauses. Updating the filter with a clean PSD estimated during speech absence indeed provides poor performance as speech resumes.

The DNN used in [19] is composed of a single long-short term memory (LSTM) layer with 512 units followed by two linear layers with rectified linear activations (ReLU), and a linear output layer with sigmoid activation. We remove the two ReLU-activated layers in our experiments, as it did not degrade the dereverberation performance, while reducing by 75 % the number of trainable parameters.

4.3. Compared algorithms

The algorithms evaluated are:

- RLS-WPE using the target PSD (*Oracle-WPE*)
- Classical RLS-WPE (*Vanilla-WPE*)[18]
- DNN-supported RLS-WPE (*DNN-WPE*) [19]
- Proposed end-to-end RLS-WPE (*E2E-WPE*)
- Proposed pretrained E2E-WPE (*E2E-WPE-p*)

The suffixes *HA* and *CI* correspond to the hearing-aided and cochlear-implanted scenarios, respectively.

4.4. Evaluation metrics

We evaluate all approaches on the described test sets. The evaluation is conducted in terms of early-to-late reverberation ratio (ELR) [27], perceptual evaluation of speech quality (PESQ), extended short-time objective intelligibility (ESTOI) [28] and signal-to-distortion ratio (SDR) [29]. The ELR computation uses a separation time of 40 ms, and is not applicable to evaluating the CI scenario since the target is the direct path only.

5. RESULTS AND DISCUSSION

We first evaluate the Oracle-WPE approach in the HA scenario, over the first 4 seconds interval and after. As indicated in Table 1, WPE performance is substantially worse when the filter is not fully initialized. In all further experiments, this initialization period is excluded from evaluation. We then compare the mentioned approaches in the HA scenario (Table 2) and the CI scenario (Table 3).

We notice that for all T_{60} and scenarios, the proposed E2E-WPE-p outperforms its DNN-WPE and Vanilla-WPE counterparts on all metrics. This shows that taking the WPE dereverberation algorithm

	0.4 → 0.6				0.6 → 0.8				0.8 → 1.0				Average			
	ELR	PESQ	ESTOI	SDR	ELR	PESQ	ESTOI	SDR	ELR	PESQ	ESTOI	SDR	ELR	PESQ	ESTOI	SDR
Unprocessed	*4.9*	*2.49*	*0.70*	*2.8*	*2.2*	*2.22*	*0.59*	*0.2*	*0.3*	*2.04*	*0.51*	*-1.6*	*2.5*	*2.25*	*0.60*	*0.5*
Oracle-WPE-HA	*11.0*	*3.19*	*0.85*	*5.8*	*7.0*	*2.77*	*0.77*	*2.8*	*4.7*	*2.52*	*0.70*	*0.9*	*7.6*	*2.83*	*0.77*	*3.2*
Vanilla-WPE	11.5	3.00	0.84	6.4	8.2	2.63	0.75	4.0	6.0	2.41	0.68	2.3	8.6	2.68	0.76	4.2
DNN-WPE-HA	11.3	3.06	0.85	6.1	7.5	2.67	0.76	3.4	5.1	2.43	0.69	1.5	8.0	2.72	0.77	3.7
E2E-WPE-HA	13.5	3.00	0.84	6.8	9.9	2.68	0.77	4.6	7.4	2.46	0.70	3.0	10.3	2.71	0.77	4.8
E2E-WPE-p-HA	**13.7**	**3.07**	**0.86**	**6.9**	**10.6**	**2.73**	**0.78**	**4.7**	**7.8**	**2.49**	**0.71**	**3.1**	**10.5**	**2.76**	**0.78**	**4.9**

Table 2. Evaluation results on the HA test set, for different T_{60} reverberation times indicated on the top row in seconds. For all metrics, the higher the better. Best performance is indicated in bold.

	0.4 → 0.6				0.6 → 0.8				0.8 → 1.0				Average			
	ELR	PESQ	ESTOI	SDR	ELR	PESQ	ESTOI	SDR	ELR	PESQ	ESTOI	SDR	ELR	PESQ	ESTOI	SDR
Unprocessed	-	*2.29*	*0.58*	*-8.8*	-	*2.05*	*0.49*	*-10.4*	-	*1.89*	*0.42*	*-11.6*	-	*2.08*	*0.50*	*-10.3*
Oracle-WPE-CI	-	*2.91*	*0.76*	*-6.3*	-	*2.57*	*0.68*	*-8.1*	-	*2.36*	*0.61*	*-9.3*	-	*2.61*	*0.68*	*-7.9*
Vanilla-WPE	-	2.71	0.72	-6.3	-	2.41	0.64	-7.6	-	2.21	0.58	-8.7	-	2.44	0.65	-7.6
DNN-WPE-CI	-	2.74	0.73	-6.7	-	2.43	0.65	-8.4	-	2.23	0.59	-9.6	-	2.47	0.66	-8.2
E2E-WPE-CI	-	2.79	0.75	**-6.0**	-	2.49	0.68	**-7.4**	-	2.28	0.62	**-8.4**	-	2.52	0.68	**-7.3**
E2E-WPE-p-CI	-	**2.83**	**0.76**	-6.2	-	**2.53**	**0.69**	-7.6	-	**2.32**	**0.63**	-8.6	-	**2.56**	**0.69**	-7.4

Table 3. Evaluation results on CI test set, for different T_{60} reverberation times indicated on the top row in seconds. For all metrics, the higher the better. Best performance is indicated in bold.

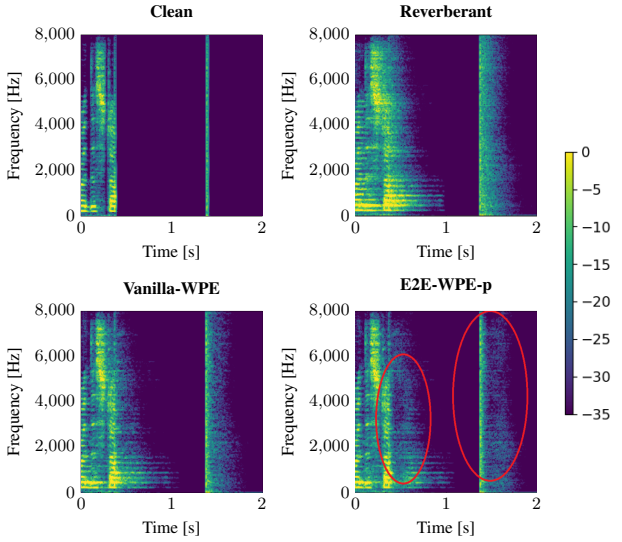


Fig. 2. Log-energy spectrograms of clean, reverberant and processed signals. Dirac impulse following an utterance. $T_{60} = 0.75$s.

into account in the DNN optimization process allows the approach to reach an improved output result, without adding any computation nor prior information at test time. We notice that on all metrics except PESQ, the E2E-WPE-p approach performs even slightly better than Oracle-WPE. Our interpretation is that through the end-to-end training procedure, the network does not try to produce an optimal PSD but rather an optimal output. Thus it implicitly modifies the probabilistic nature of the parameter $\lambda_{t,f}$, which then plays the role of a regularizer in (3) rather than that of a variance. Possible explanations are that it either relaxes the Gaussian assumption on the anechoic speech s [9] or corrects the bias in estimating the time-varying PSD via the periodogram in (6). As can be seen in Table 2, using a pretrained DNN significantly helps improving the performance.

Although a filter length of $K = 10$ frames and a delay of $\Delta = 5$ frames (in the HA scenario) only permits to fully cancel reverberation up to 120 ms, all approaches achieve significant dereverberation for T_{60} up to 1.0s. Indeed, the reverberation energy decaying approximately exponentially [1], the major part of it resides in the beginning of the reverberation tail. Therefore, although we perceive remains of late reverberation, the objective results are good, especially for the ELR metric which highly reflects this phenomenon.

This contrast between objective improvement and residual reverberation is emphasized with the proposed E2E-WPE(-p) approaches. This is shown in Figure 2 where an utterance is used to initialize the DNN and WPE statistics and a Dirac impulse is added following 1 second of silence. We notice that the speech contains less short and moderate reverberant energy, yielding a good ELR improvement although some residual late reverberation is present. This is also in line with our informal listening experiments. With the DNN-WPE and Vanilla-WPE approaches, the late reverberation is less identifiable as it is obfuscated by the energy remaining in the short and moderate reverberation through the time-masking phenomenon.

Several approaches to further improve the results may be considered, for instance noise reduction post-processing. As residual late reverberation is perceptually close to noise, it would potentially be a good target for such methods. This is preferred to increasing the prediction filter length of our approach, which results in industrious training while still being unable to cancel very long reverberation.

6. CONCLUSION

We proposed an end-to-end training procedure of the DNN-supported WPE dereverberation algorithm based on [19]. The traditional signal processing computations were included into the training of the neural network estimating the anechoic speech PSD. This allowed for specialized training with respect to needs of different listener categories, by letting the network learn customized WPE parameters and targets. Results show that this training procedure improved the dereverberation performance without extra computational cost. The approach suppressed most of the reverberation energy immediately following the early reflections, and could be combined with subsequent post-filtering for removing residual late reverberation.

7. REFERENCES

[1] H. Kuttruff, "Room acoustics," *CRC Press*, 2016.

[2] P. Naylor and N. Gaubitch, *Speech Dereverberation*, vol. 59. 01 2011.

[3] E. Habets, *Single- and Multi-Microphone Speech Dereverberation Using Spectral Enhancement*. PhD thesis, 01 2007.

[4] A. Kuklasiński, S. Doclo, T. Gerkmann, S. Holdt Jensen, and J. Jensen, "Multi-channel PSD estimators for speech dereverberation - a theoretical and experimental comparison," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, pp. 91–95, 2015.

[5] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukic, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Proc.*, vol. 2015, pp. 1–12, 2015.

[6] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *The Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, 1977.

[7] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 23, no. 6, pp. 1006–1018, 2015.

[8] T. Gerkmann, "Cepstral weighting for speech dereverberation without musical noise," in *2011 19th European Signal Processing Conference*, pp. 2309–2313, 2011.

[9] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *ICASSP 2008*, pp. 85–88, 2008.

[10] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 23, no. 9, pp. 1509–1520, 2015.

[11] A. Warzybok, J. Rennies, S. D. T. Brand, and B. Kollmeier, "Effects of spatial and temporal integration of a single early reflection on speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 2947–2952, 2003.

[12] H. S. J. S. Bradley and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 2947–2952, 2003.

[13] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 19, no. 1, pp. 69–84, 2011.

[14] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, pp. 31–35, 2018.

[15] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation," in *ISCA Interspeech*, 2017.

[16] B. Schwartz, S. Gannot, and E. A. P. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 23, no. 2, pp. 394–406, 2015.

[17] S. Braun and E. A. P. Habets, "Online dereverberation for dynamic scenarios using a Kalman filter with an autoregressive model," *IEEE Signal Proc. Letters*, vol. 23, no. 12, pp. 1741–1745, 2016.

[18] J. Caroselli, I. Shafran, A. Narayanan, and R. Rose, "Adaptive multichannel dereverberation for automatic speech recognition," in *ISCA Interspeech*, 2017.

[19] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Frame-online DNN-WPE dereverberation," *IWAENC*, pp. 466–470, 2018.

[20] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "Mimo-speech: End-to-end multi-channel multi-speaker speech recognition," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 237–244, 2019.

[21] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Proc.*, vol. 11, no. 8, pp. 1274–1288, 2017.

[22] W. Zhang, C. Boeddeker, S. Watanabe, T. Nakatani, M. Delcroix, K. Kinoshita, T. Ochiai, N. Kamo, R. Haeb-Umbach, and Y. Qian, "End-to-end dereverberation, beamforming, and speech recognition with improved numerical stability and advanced frontend," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2021.

[23] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Joint optimization of neural network-based WPE dereverberation and acoustic model for robust online ASR," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2019.

[24] Y. Hu and K. Kokkinakis, "Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners," *The Journal of the Acoustical Society of America*, vol. 135, pp. EL22–8, 01 2014.

[25] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "Whamr!: Noisy and reverberant single-channel speech separation," 2020.

[26] T. Wendt, S. Van De Par, and S. D. Ewert, "A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation," *Journal of the Audio Engineering Society*, vol. 62, pp. 748–766, november 2014.

[27] G. Carbajal, R. Serizel, E. Vincent, and E. Humbert, "Joint NN-supported multichannel reduction of acoustic echo, reverberation and noise," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 28, pp. 2158–2173, 2020.

[28] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 24, pp. 1–1, 11 2016.

[29] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.

A.2 Neural Network-augmented Kalman Filtering for Robust Online Speech Dereverberation in Noisy Reverberant Environments [P3]

Abstract

In this paper, a neural network-augmented algorithm for noise-robust online dereverberation with a Kalman filtering variant of the weighted prediction error (WPE) method is proposed. The filter stochastic variations are predicted by a deep neural network (DNN) trained end-to-end using the filter residual error and signal characteristics. The presented framework allows for robust dereverberation on a single-channel noisy reverberant dataset similar to WHAMR!. The Kalman filtering WPE introduces distortions in the enhanced signal when predicting the filter variations from the residual error only, if the target speech power spectral density is not perfectly known and the observation is noisy. The proposed approach avoids these distortions by correcting the filter variations estimation in a data-driven way, increasing the robustness of the method to noisy scenarios. Furthermore, it yields a strong dereverberation and denoising performance compared to a DNN-supported recursive least squares variant of WPE, especially for highly noisy inputs.

Reference

Jean-Marie Lemercier, Joachim Thiemann, Raphael Koning and Timo Gerkmann "Neural Network-augmented Kalman Filtering for Robust Online Speech Dereverberation in Noisy Reverberant Environments", *ISCA Interspeech*, 2022. DOI: 10.21437/Interspeech.2022-11337

Copyright Notice

Authors' Contributions

Jean-Marie Lemercier is the first author of this publication. He implemented all algorithms, trained the neural networks used in the paper, conducted the experimental validation, and wrote the manuscript. Timo Gerkmann brought insights on the experimental validation, mathematical derivations, and also thoroughly reviewed the paper. Joachim Thiemann and Raphael Koning brought their feedback on all methods through discussions, they also helped with reviewing the manuscript.

Neural Network-augmented Kalman Filtering for Robust Online Speech Dereverberation in Noisy Reverberant Environments

Jean-Marie Lemercier[⋆]*, Joachim Thiemann*[†]*, Raphael Koning*[†]*, Timo Gerkmann*[⋆]

[⋆]Signal Processing (SP), Universität Hamburg, Germany
[†]Advanced Bionics, Hanover, Germany

{firstname.lastname}@uni-hamburg.de, {firstname.lastname}@advancedbionics.com

Abstract

In this paper, a neural network-augmented algorithm for noise-robust online dereverberation with a Kalman filtering variant of the weighted prediction error (WPE) method is proposed. The filter stochastic variations are predicted by a deep neural network (DNN) trained end-to-end using the filter residual error and signal characteristics. The presented framework allows for robust dereverberation on a single-channel noisy reverberant dataset similar to WHAMR!.

The Kalman filtering WPE introduces distortions in the enhanced signal when predicting the filter variations from the residual error only, if the target speech power spectral density is not perfectly known and the observation is noisy. The proposed approach avoids these distortions by correcting the filter variations estimation in a data-driven way, increasing the robustness of the method to noisy scenarios. Furthermore, it yields a strong dereverberation and denoising performance compared to a DNN-supported recursive least squares variant of WPE, especially for highly noisy inputs.

Index Terms: dereverberation, kalman filtering, adaptive processing, neural network, end-to-end training

1. Introduction

Communication and hearing devices require modules aiming at suppressing undesired parts of the signal to improve the speech characteristics. Amongst these is reverberation caused by room acoustics, where late reflections particularly degrade the speech quality and intelligibility [1]. Presence of additional background noise and interfering speakers further worsens the ability to clearly perceive target speech.

In complement to traditional single-channel schemes, many multi-channel algorithms leveraging spatial and spectral information were proposed for enhancement of noisy reverberant speech. Traditional approaches include beamforming [2, 3, 4], possibly combined with spectral enhancement [5], coherence-weighting [6, 7], and multi-channel linear prediction (MCLP) based approaches such as the well-known weighted prediction error (WPE) algorithm [8, 9]. WPE computes an auto-regressive multi-channel filter in the short-time spectrum and applies it to a delayed group of reverberant speech frames. It requires an estimate of the target speech power spectral density (PSD), estimated either by statistical models [8, 9] or deep neural networks (DNNs) [10, 11].

In order to cope with real-time requirements and changing acoustics, online adaptive dereverberation methods derived from MCLP approaches were introduced. These methods are based on either Kalman filtering (KF) [12, 13, 14, 15, 16] or on a recursive least squares (RLS) adapted WPE, which can be seen a special case of KF [17, 18, 11, 19]. Online convolutional beamformers performing joint dereverberation and denoising based on either RLS or KF were proposed in [20, 21, 22, 15, 16].

Kalman filtering WPE (KF-WPE) is a particularly interesting framework for adaptive dereveberation in noisy and dynamic environments. It updates the filter in a much faster and more flexible way than its RLS counterpart, due to a Gauss-Markov model of the filter transition dynamics [23]. However, it is known that the Kalman filter can be particularly sensitive to estimation errors in the unobservable state space model parameters [24]. In KF-WPE, these parameters include the target speech PSD and the filter transition covariance.

The contributions of this work are threefold. First, we introduce a low-complexity variant of KF-WPE based on [13] and a scaled identity model of the speech covariance matrix. Secondly, we evaluate the performance and robustness of this KF-WPE variant in noisy reverberant environments. Finally, we present an estimation strategy of both the target speech PSD and the filter transition covariance based on DNNs. In this strategy, a first DNN estimates the target speech PSD from the noisy reverberant mixture. A second DNN then estimates the filter transition covariance from the filter residual error and signal characteristics, and is trained with an end-to-end criterion.

This framework enables a robust estimation of the filter transition covariance under target PSD uncertainty. The resulting DNN-augmented KF-WPE performs stronger dereverberation and denoising than RLS-WPE, with less distortions than traditional KF-WPE — all approaches using the same DNN-supported target PSD estimator. The approach is robust to noisy observations and even removes a lot of environmental noise, although noise is not accounted for in the signal model.

The rest of this paper is organized as follows. In Section 2, the low-complexity variant of KF-WPE is summarized. Section 3 presents the DNN-supported strategy to predict the WPE filter transition covariance and the target speech PSD. In Section 4, we describe the experimental setup and training strategy. Results are presented and discussed in Section 5.

2. Kalman filtering adapted WPE dereverberation

2.1. Signal model

In the short-time Fourier transform (STFT) domain using the subband-filtering approximation [8], the noisy reverberant speech $\boldsymbol{x} = [x^{(1)}, \dots, x^{(D)}] \in \mathbb{C}^D$ is obtained at the microphone array by convolution of the anechoic speech s and the room impulse responses (RIRs) $\boldsymbol{H} \in \mathbb{C}^{D \times D}$ with length N:

This work has been funded by the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380. The authors are responsible for the content of this paper.

$$\boldsymbol{x}_{t,f} = \sum_{\tau=0}^{N} \boldsymbol{H}_{\tau,f} s_{t-\tau,f} + \boldsymbol{n}_{t,f} \tag{1}$$

$$= \boldsymbol{d}_{t,f} + \boldsymbol{e}_{t,f} + \boldsymbol{r}_{t,f} + \boldsymbol{n}_{t,f}, \tag{2}$$

where t denotes the time frame index and f the frequency bin, which we will drop when not needed. $\boldsymbol{d}$ denotes the direct path, $\boldsymbol{e}$ the early reflections component, $\boldsymbol{r}$ the late reverberation and $\boldsymbol{n}$ an error term comprising modelling errors and environmental noise. The early reflections $\boldsymbol{e}$ were shown to contribute to speech quality and intelligibility for normal and hearing-aided listeners [25]. The dereverberation objective is therefore to retrieve $\boldsymbol{\nu} = \boldsymbol{d} + \boldsymbol{e}$.

However, early reflections may be detrimental for some people, e.g. cochlear-implant users, particularly in highly-reverberant scenarios [26]. Accordingly, the dereverberation objective can be adjusted to different listener categories [19].

2.2. WPE Dereverberation

As in [8], the anechoic speech s is modelled in the STFT domain with a zero-mean time-varying Gaussian model. It follows from (1) that the target speech also is a zero-mean time-varying Gaussian process with time-frequency dependent covariance:

$$\boldsymbol{\nu}_{t,f} \sim \mathcal{N}_{\mathbb{C}}(0; \boldsymbol{\Phi}^{(\nu)}_{t,f}) \tag{3}$$

The WPE algorithm [8] uses an auto-regressive model to approximate the late reverberation $\boldsymbol{r}$. A multi-channel filter $\boldsymbol{G} \in \mathbb{C}^{D^2K}$ with K taps is estimated, aiming at representing the inverse of the late tail of the RIRs $\boldsymbol{H}$. The target $\boldsymbol{\nu}$ is then obtained through linear prediction, with a delay Δ avoiding undesired short-time speech cancellations, which also leads to preserving parts of the early reflections. We omit the frequency index f in the following as computations are performed in each frequency band independently. By disregarding the error term $\boldsymbol{n}$ in (1) in noiseless scenarios we obtain:

$$\boldsymbol{\nu}^{(\mathrm{WPE})}_t = \boldsymbol{x}_t - \boldsymbol{G}^H (\boldsymbol{I}_D \otimes \boldsymbol{X}_{t-\Delta}), \tag{4}$$

where $\boldsymbol{X}_{t-\Delta} = \left[\boldsymbol{x}^T_{t-\Delta}, \ldots, \boldsymbol{x}^T_{t-\Delta-K+1}\right]^T \in \mathbb{C}^{DK}$ and $\otimes$ is the Kroenecker product.

2.3. Kalman Filtering WPE Dereverberation

In order to obtain an adaptive and real-time capable approach, a KF variant of WPE was proposed in [13], where the WPE filter $\boldsymbol{G}$ is recursively updated with a Markov model:

$$\boldsymbol{G}_t = \boldsymbol{G}_{t-1} + \boldsymbol{q}_t, \tag{5}$$

where $\boldsymbol{q}$ is the filter transition stochastic noise following a zero-mean Gaussian process $\boldsymbol{q}_t \sim \mathcal{N}_{\mathbb{C}}(0; \boldsymbol{Q}_t)$.

A Gaussian-Markov state-space model is formed by the transition equation (5) and the observation model (4), where $\boldsymbol{G}$ is the state, $\boldsymbol{x}$ the observation and $\boldsymbol{\nu}^{(\mathrm{WPE})}$ the *observation noise*. Given these and some independency assumptions described in [13], the Kalman filter $\widehat{\boldsymbol{G}}$ is an optimal recursive estimator with respect to the mean-squared error criterion:

$$\arg\min_{\boldsymbol{G}} \mathbb{E}_t\{\, ||\boldsymbol{G}_t - \widehat{\boldsymbol{G}}_t||_2^2 \,\}, \tag{6}$$

Similarly to [8, 17], we make the further assumption that the target speech $\boldsymbol{\nu}$ is modelled identically and independently at each microphone, thus making the target speech covariance a scaled identity matrix characterized by PSD λ:

$$\boldsymbol{\Phi}^{(\nu)}_t = \lambda_t \boldsymbol{I}_D \tag{7}$$

The WPE filter $\boldsymbol{G}^{(d)} \in \mathbb{C}^{DK}$ is then estimated and applied for each channel d independently, which considerably curbs the algorithmic complexity in a multi-channel setting.

We define the filter error covariance matrix as:

$$\boldsymbol{\Phi}^{(\epsilon)} = \mathbb{E}_{t,d}\{\, [\boldsymbol{G}^{(d)}_t - \widehat{\boldsymbol{G}}^{(d)}_t][\boldsymbol{G}^{(d)}_t - \widehat{\boldsymbol{G}}^{(d)}_t]^H \,\} \tag{8}$$

The corresponding derivations are adapted from [13]:

$$\boldsymbol{\Phi}^{(\epsilon)}_{t|t-1} = \boldsymbol{\Phi}^{(\epsilon)}_{t-1|t-1} + \boldsymbol{Q}_{t-1}, \tag{9}$$

$$\boldsymbol{K}_t = \frac{\boldsymbol{\Phi}^{(\epsilon)}_{t|t-1} \boldsymbol{X}_{t-\Delta}}{\lambda_t + \boldsymbol{X}^H_{t-\Delta} \boldsymbol{\Phi}^{(\epsilon)}_{t|t-1} \boldsymbol{X}_{t-\Delta}}, \tag{10}$$

$$\boldsymbol{\Phi}^{(\epsilon)}_{t|t} = \boldsymbol{\Phi}^{(\epsilon)}_{t|t-1} - \boldsymbol{K}_t \boldsymbol{X}^T_{t-\Delta} \boldsymbol{\Phi}^{(\epsilon)}_{t|t-1}, \tag{11}$$

$$\widehat{\boldsymbol{G}}^{(d)}_t = \widehat{\boldsymbol{G}}^{(d)}_{t-1} + \boldsymbol{K}_t \big(x^{(d)}_t - (\widehat{\boldsymbol{G}}^{(d)}_{t-1})^H \boldsymbol{X}_{t-\Delta,f}\big)^H, \tag{12}$$

$$\widehat{\nu}^{(d)}_t = x^{(d)}_t - (\widehat{\boldsymbol{G}}^{(d)}_t)^H \boldsymbol{X}_{t-\Delta,f}. \tag{13}$$

where $\boldsymbol{K} \in \mathbb{C}^{DK}$ is the Kalman gain.

2.4. RLS-WPE dereverberation

A RLS version of WPE was introduced in [17] and is equivalent to KF-WPE for the static case, i.e. when $\boldsymbol{q} = 0$. The target speech covariance is replaced by a scaled identity matrix as in (7). The filter error covariance is thus equivalent to the inverse of the weighted covariance of the reverberant buffer:

$$\boldsymbol{\Phi}^{(\epsilon)}_{\mathrm{RLS}} = \left[\mathbb{E}_t \left\{ \frac{\boldsymbol{X}_{t-\Delta} \boldsymbol{X}^H_{t-\Delta}}{\lambda_t} \right\}\right]^{-1}. \tag{14}$$

The computations are equivalent to (9)-(12), only that the filter error covariance is updated recursively with forgetting factor α [23], replacing (9) with:

$$\boldsymbol{\Phi}^{(\epsilon)}_{\mathrm{RLS},t|t-1} = \frac{1}{\alpha} \boldsymbol{\Phi}^{(\epsilon)}_{\mathrm{RLS},t-1|t-1}. \tag{15}$$

Processing order is then (15)→(10)→(11)→(12)→(13).

3. Filter transition covariance estimation

3.1. Covariance model

The filter transition covariance $\boldsymbol{Q}$ is an unknown parameter which must be estimated at each step to update the filter error covariance $\boldsymbol{\Phi}^{(\epsilon)}$ in (9). As in [13] we model each filter tap transition identically and independently, which results in $\boldsymbol{Q}$ being an identity matrix scaled by the *filter transition power* $\phi^{(q)}$:

$$\boldsymbol{Q}_t = \phi^{(q)}_t \boldsymbol{I}_{DK}. \tag{16}$$

We furthermore define the *filter residual error* e as:

$$e_t = \mathbb{E}_d\{\, ||\boldsymbol{G}^{(d)}_t - \boldsymbol{G}^{(d)}_{t-1}||_2^2 \,\}. \tag{17}$$

In [13], the filter transition power is simply modelled by adding a small fixed parameter η to the scaled residual error e, in order to force a permanent adaptation even if the filter did not vary at the previous time step:

$$\phi^{(q)}_t = \frac{1}{DK} e_t + \eta \tag{18}$$

3.2. DNN estimation

As we will show in Section 5, using the filter residual error e to model the filter transition power $\phi^{(q)}$ is a straightforward model which yields excellent results if the oracle target PSD is available. However, if the PSD estimation is flawed due to noisy observations and limited prediction power, using the same filter transition power model (17) introduces problematic speech distortions. In particular, the bias η should be adapted as a function of the input signal-to-noise ratio (SNR), decreasing the filter transition power if the observation is too uncertain.

Rather than using a statistical approach based on input SNR analysis, we propose here a DNN strategy to infer at each frame the filter transition power $\phi^{(q)}$ directly from data, using an end-to-end criterion to optimize the DNN parameters. We argue that this increases the robustness of KF-WPE to erroneous PSD estimation. A related concept is reported in [27], where a DNN is used to learn the step size, i.e. the Kalman gain, of a block-wise adaptive system identification algorithm. Our approach is however distinct as (i)- we perform frame-wise adaptive dereverberation in noisy environments, (ii)- we do not use a single DNN inferring both step size parameters, but use separated networks with distinct inputs and training strategies; and (iii)- the criterion optimizes the DNNs with respect to the estimated signal, and not the filter error, which is not available in this task.

The DNN—called here VarNet—takes as its input at every step t a vector containing the channel-averaged reverberant speech periodogram $|\bar{x}_t|^2 = \frac{1}{D}\sum_{d=1}^{D}|x_t^{(d)}|^2\}$, the target speech PSD estimate $\widehat{\lambda}_t$ and the filter residual error e_t. The filter transition power $\phi_t^{(q)}$ is then obtained on a model inspired by (18), where the VarNet positive real-valued output mask $\mathcal{M}_{\sqcup}^{(\eta)}$ is multiplied by a maximal bias value $\eta_{\max}$ before being added to the scaled filter residual error e:

$$\phi_t^{(q)} = \frac{1}{DK}e_t + \eta_{\max} \odot \mathcal{M}_t^{(\eta)}(|\bar{x}_t|^2, \widehat{\lambda}_t, e_t) \qquad (19)$$

The target speech PSD estimate $\widehat{\lambda}_t$ is obtained from the channel-averaged magnitude $|\bar{x}_t| = \frac{1}{D}\sum_{d=1}^{D}|x_t^{(d)}|\}$ by a preceding DNN—denoted as $\mathrm{MaskNet}$—as in [11, 19]:

$$\widehat{\lambda}_t = \left(\mathcal{M}_t^{(\nu)} \odot |\bar{x}_t|\right)^2, \qquad (20)$$

with $\mathcal{M}_t^{(\nu)}$ being the estimated real-valued positive mask.

3.3. Training strategy

The $\mathrm{MaskNet}$ is pre-trained with a mask-based objective:

$$\mathcal{L}^{(\mathrm{pre})} = \mathcal{L}(\mathcal{M}^{(\nu)} \odot |\boldsymbol{x}|, |\boldsymbol{\nu}|), \qquad (21)$$

with $\mathcal{L}$ being the chosen loss function. The parameters of $\mathrm{MaskNet}$ are then frozen, and the VarNet is trained with the end-to-end criterion:

$$\mathcal{L}^{(\mathrm{end})} = \mathcal{L}(|\widehat{\boldsymbol{\nu}}|, |\boldsymbol{\nu}|). \qquad (22)$$

Schematics of the algorithm are displayed in Figure 1.

4. Experimental setup

4.1. Dataset generation

The data generation method resembles that of the WHAMR! dataset [28]. We concatenate anechoic speech utterances from the WSJ0 dataset belonging to the same speaker, and construct

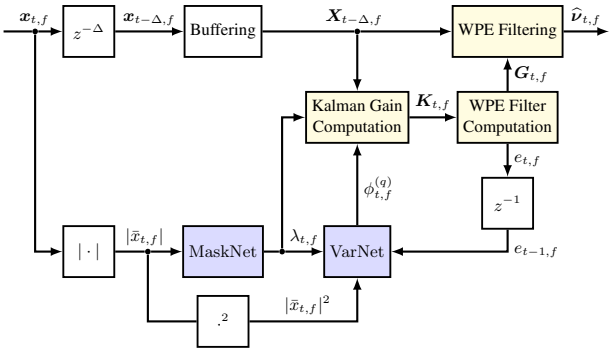


Figure 1: *DNN-supported Kalman-filtering adapted dereverberation. Blue blocks refer to trainable DNN layers. Yellow blocks represents adaptive statistical signal processing. $z^{-\tau}$ blocks implement a time delay of τ STFT frames.*

sequences of approximately 20 seconds. These sequences are convolved with 2-channel RIRs generated with the RAZR engine [29] and randomly picked. Each RIR is generated by uniformly sampling room acoustics parameters as in [28] and a T_{60} reverberation time between 0.4 and 1.0 seconds. Head-related transfer function (HRTF) auralization is performed in the RAZR engine, using a KEMAR dummy head response from the MMHR-HRTF database [30]. Finally, 2-channel noise from the WHAM! recorded noise dataset [31] is added to the reverberant mixture with a SNR—relative to the reverberant signal—uniformly sampled between -5 and 25 dB .

Because of its adaptive nature, the online variants of WPE have an initialization time L_i (typically 4s with the hyperparameters described below). The utterances are therefore cut into segments of length L_i, and the first segment is not used by the end-to-end training procedure, as described in [19, 32].

Target datapoints are obtained by convolving the first 40 ms of the RIR with the dry speech utterance, simulating the direct path and early reflections, which are beneficial for hearing-aided and normal-hearing listeners [25].

400,100, 60 RIRs and 20000, 5000, 3000 clean utterances and noisy excerpts are used for training, validation and testing respectively. Ultimately, each training set consists of approximately 55 hours of speech data sampled at 16 kHz.

4.2. Hyperparameter settings

All approaches are trained with the Adam optimizer using a learning rate of 10^{-4} and mini-batches of 128 segments. The $\mathrm{MaskNet}$ is pre-trained for 300 epochs, and the VarNet is trained for 100 epochs All networks are optimized with respect to a L^1 loss on the spectrogram magnitude. DNN inputs are standardized using the mean and variance of the noisy reverberant distribution approximated by the training set.

The STFT uses a square-rooted Hann window of 32 ms and a 75 % overlap, which yields $F = 257$ frequency bins with the corresponding sampling frequency. The WPE filter length is set to $K = 10$ STFT frames (i.e. 80 ms), the number of channels to $D = 2$, the prediction delay to $\Delta = 5$ frames (i.e. 40 ms). The prediction delay value is picked as it experimentally provides optimal evaluation metrics when using the oracle PSD, and matches the target set in the previous subsection. When not learnt by VarNet, the filter transition bias is set to $\eta = -35$ dB as in [13], and when it is learnt, the maximal bias is set to $\eta_{\max} = -30$dB.

The $\mathrm{MaskNet}$ structure is the same as used in [19, 32], that is, a single long short-term memory (LSTM) layer with 512

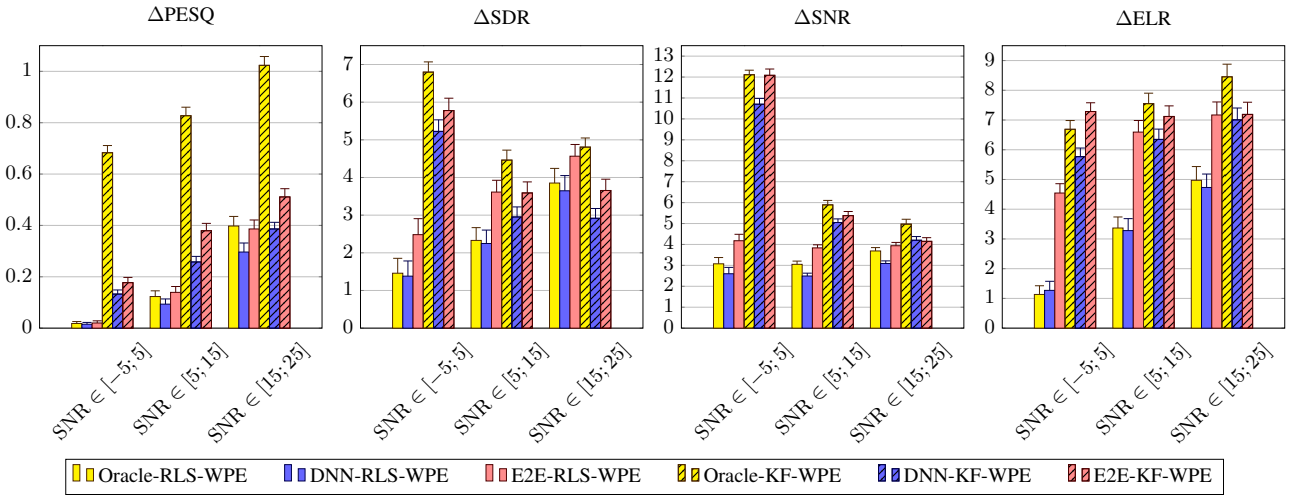


Figure 2: *Improvements upon noisy reverberant signals. All metrics except PESQ are in* dB. *Input SNRs are indicated in* dB.

units followed by a linear output layer with sigmoid activation. The VarNet uses a first linear layer with F output units to fuse the input modalities. A single LSTM with 512 hidden units is then used to model the sequence dynamics, and is followed by a linear layer with F units and sigmoid activation.

We estimate the number of MAC operations per second of our proposed algorithm to 31.4 GMAC·s^{-1} at 16 kHz. This setting allows to obtain a real-time factor—defined as the ratio between the time needed to process an utterance and the length of the utterance—below 0.15 with all computations performed on Intel(R) Core(TM) i7-9800X CPU.

4.3. Evaluation metrics

We evaluate our algorithms on the noisy reverberant test set with the Perceptual Evaluation of Listening Quality (PESQ) [33], output signal-to-distortion ratio (SDR) and SNR [34], as well as the early-to-late reverberation ratio (ELR) [1, 35, 19].

5. Results and discussion

5.1. Compared approaches

We evaluate the RLS-WPE and KF-WPE using the oracle target PSD λ (**Oracle-RLS-WPE** and **Oracle-KF-WPE**). We then compare the RLS-WPE approach using the DNN-estimated PSD (**DNN-RLS-WPE**, [11]) and KF-WPE using the DNN-estimated PSD and the filter transition power model (17) (**DNN-KF-WPE**, [13]). Finally, we evaluate the proposed approach using KF-WPE with DNN-estimated PSD and filter transition power (19) (**E2E-KF-WPE**). We also include a RLS-WPE algorithm where the target PSD λ is estimated by a DNN pre-trained using (21) and fine-tuned end-to-end with (22) (**E2E-RLS-WPE**, [19]). Results are displayed in Figure 2.

5.2. Oracle experiments

We first notice that KF-WPE is largely superior to RLS-WPE, if the oracle target PSD λ is used. In particular, the SNR and ELR scores of Oracle-KF-WPE indicate that most of the reverberation was removed, although the length of the WPE filter only covers 80 ms, which is largely inferior to the T_{60} times used ($0.4 - 1.0$ s). It is also able to remove noise very efficiently, although no denoising mechanism is specified in the signal model.

5.3. DNN-assisted frameworks

If the target PSD is estimated by MaskNet without learning the transition power $\phi^{(q)}$, KF-WPE yields agressive dereverberation and denoising performance, thus introducing distortions in the signal. This is confirmed by the high SNR and the low PESQ and SDR scores of DNN-KF-WPE.

5.4. End-to-end frameworks

The proposed E2E-KF-WPE provides superior PESQ and SDR compared to DNN-KF-WPE, which shows that it is able to circumvent the degrading behaviour of the latter approach. Learning the filter transition power $\phi^{(q)}$ helps controlling the adaptation speed as a function of the noise and reverberation condition, thus yielding higher robustness to the estimation errors from MaskNet. Also, E2E-KF-WPE exhibits high SNR and ELR scores compared to its RLS-WPE counterparts. This indicates that E2E-KF-WPE is able to significantly improve the dereverberation and denoising performance by using Kalman filtering, especially for low input SNRs.

We notice in our experiments that the learnt $\phi^{(q)}$ increases with the T_{60} time and decreases with the input SNR, therefore accelerating the adaptation speed as the adversity of the condition intensifies. We hypothesize that this trend is learnt to compensate for the original model mismatch caused by the WPE filter being too short in comparison to the T_{60} time on the one hand, and to WPE being noise-agnostic on the other hand.

6. Conclusion

We presented a DNN-augmented Kalman filtering framework for robust online dereverberation in noisy reverberant environments. Through end-to-end optimization, the DNN estimating the filter transition power is able to control the WPE filter adaptation speed with respect to the noise and reverberation condition. Our algorithm thus avoids degrading the target signal compared to the original Kalman filtering WPE, while exhibiting stronger dereverberation and denoising power than its RLS-based counterparts. Future work will be dedicated to inspecting adaptation mechanisms for dynamic acoustics and to the inclusion of suitable denoising approaches in the present framework.

7. References

[1] P. Naylor and N. Gaubitch, "Speech dereverberation," *Noise Control Engineering Journal*, vol. 59, 01 2011.

[2] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, *Speech Distortion Weighted Multichannel Wiener Filtering Techniques for Noise Reduction*. Springer Berlin Heidelberg, 2005, pp. 199–228.

[3] A. Kuklasiński, S. Doclo, T. Gerkmann, S. Holdt Jensen, and J. Jensen, "Multi-channel PSD estimators for speech dereverberation - a theoretical and experimental comparison," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2015, pp. 91–95.

[4] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation with probabilistic reverberation priors," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 24, no. 12, pp. 2453–2465, Sep. 2016.

[5] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukic, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Proc.*, vol. 2015, pp. 1–12, 2015.

[6] T. Gerkmann, "Cepstral weighting for speech dereverberation without musical noise," in *Proc. Euro. Signal Proc. Conf. (EUSIPCO)*, 2011, pp. 2309–2313.

[7] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 23, no. 6, pp. 1006–1018, 2015.

[8] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2008, pp. 85–88.

[9] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 23, no. 9, pp. 1509–1520, 2015.

[10] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation," in *ISCA Interspeech*, 2017.

[11] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Frame-online DNN-WPE dereverberation," *IWAENC*, pp. 466–470, 2018.

[12] B. Schwartz, S. Gannot, and E. A. P. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 23, no. 2, pp. 394–406, 2015.

[13] S. Braun and E. A. P. Habets, "Online dereverberation for dynamic scenarios using a Kalman filter with an autoregressive model," *IEEE Signal Proc. Letters*, vol. 23, no. 12, pp. 1741–1745, 2016.

[14] ——, "Linear prediction-based online dereverberation and noise reduction using alternating Kalman filters," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 26, no. 6, pp. 1119–1129, 2018.

[15] S. Hashemgeloogerdi and S. Braun, "Joint beamforming and reverberation cancellation using a constrained Kalman filter with multichannel linear prediction," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2020, pp. 481–485.

[16] S. Braun and I. Tashev, "Low complexity online convolutional beamforming," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 136–140.

[17] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2009, pp. 3733–3736.

[18] J. Caroselli, I. Shafran, A. Narayanan, and R. Rose, "Adaptive multichannel dereverberation for automatic speech recognition," in *ISCA Interspeech*, 2017.

[19] J.-M. Lemercier, J. Thiemann, R. Konig, and T. Gerkmann, "Customizable end-to-end optimization of online neural network-supported dereverberation for hearing devices," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2022.

[20] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Proc. Letters*, vol. 26, no. 6, pp. 903–907, 2019.

[21] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone speech dereverberation, interfering speech cancellation, and noise reduction," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 28, pp. 740–754, 2020.

[22] T. Dietzen, S. Doclo, A. Spriet, W. Tirry, M. Moonen, and T. van Waterschoot, "Low-complexity Kalman filter for multi-channel linear-prediction-based blind speech dereverberation," in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, 2017, pp. 284–288.

[23] S. Haykin, "Kalman filtering and neural networks," *John Wiley & Sons, Ltd*, pp. 1–21, 2001.

[24] B. D. O. Anderson and J. B. Moore, "Optimal filtering," *Prentice Hall*, pp. 129–135, 1979.

[25] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 2947–2952, 2003.

[26] Y. Hu and K. Kokkinakis, "Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners," *The Journal of the Acoustical Society of America*, vol. 135, pp. EL22–8, 01 2014.

[27] T. Haubner, A. Brendel, and W. Kellermann, "End-to-end deep learning-based adaptation control for frequency-domain adaptive system identification," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2022.

[28] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "Whamr!: Noisy and reverberant single-channel speech separation," 2020.

[29] T. Wendt, S. Van De Par, and S. D. Ewert, "A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation," *Journal of the Audio Engineering Society*, vol. 62, no. 11, pp. 748–766, november 2014.

[30] J. Thiemann and S. van de Pars, "A multiple model high-resolution head-related impulse response database for aided and unaided ears," *EURASIP Journal on Advances in Signal Processing*, 2019.

[31] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," *CoRR*, 2019.

[32] J.-M. Lemercier, J. Thiemann, R. Konig, and T. Gerkmann, "End-to-end optimization of online neural network-supported two-stage dereverberation for hearing devices," 2022.

[33] A. W. Rix, J. G. Beerends, M. Hollier, and A. P. 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)*, vol. 2, 2001, pp. 749–752 vol.2.

[34] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[35] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 19, no. 1, pp. 69–84, 2011.

A.3 Speech Enhancement and Dereverberation With Diffusion-Based Generative Models [P6]

Abstract

In this work, we build upon our previous publication and use diffusion-based generative models for speech enhancement. We present a detailed overview of the diffusion process that is based on a stochastic differential equation and delve into an extensive theoretical examination of its implications. Opposed to usual conditional generation tasks, we do not start the reverse process from pure Gaussian noise but from a mixture of noisy speech and Gaussian noise. This matches our forward process which moves from clean speech to noisy speech by including a drift term. We show that this procedure enables using only 30 diffusion steps to generate high-quality clean speech estimates. By adapting the network architecture, we are able to significantly improve the speech enhancement performance, indicating that the network, rather than the formalism, was the main limitation of our original approach. In an extensive cross-dataset evaluation, we show that the improved method can compete with recent discriminative models and achieves better generalization when evaluating on a different corpus than used for training. We complement the results with an instrumental evaluation using real-world noisy recordings and a listening experiment, in which our proposed method is rated best. Examining different sampler configurations for solving the reverse process allows us to balance the performance and computational speed of the proposed method. Moreover, we show that the proposed method is also suitable for dereverberation and thus not limited to additive background noise removal.

Reference

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay and Timo Gerkmann, "Speech Enhancement and Dereverberation With Diffusion-Based Generative Models", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 31, pp. 2351-2364, 2023, DOI: 10.1109/TASLP.2023.3285241

Copyright Notice

Authors' Contributions

Julius Richter and Simon Welker designed together the original SGMSE algorithm [221] and Julius Richter extended it to the SGMSE+ algorithm paper presented here. Julius Richter performed the denoising experiments and wrote most of the manuscript. Simon Welker helped with writing and revising the manuscript and preparing figures. He brought discussions on the method and performed experiments. Jean-Marie Lemercier is the third author of this publication, he handled the dereverberation experiments (Sections V-A-3, V-F-7, V-F-8, and VI-B) and helped reviewing the manuscript. Bulong Lay helped with the denoising experiments by training the ConvTasNet baseline (Section V-F-6) and helped reviewing the manuscript. Timo Gerkmann brought insights on the experimental validation, mathematical derivations, and reviewed the manuscript.

Speech Enhancement and Dereverberation with Diffusion-based Generative Models

Julius Richter, *Student Member, IEEE*, Simon Welker, *Student Member, IEEE*, Jean-Marie Lemercier, *Student Member, IEEE*, Bunlong Lay, Timo Gerkmann, *Senior Member, IEEE*

***Abstract*—In this work, we build upon our previous publication and use diffusion-based generative models for speech enhancement. We present a detailed overview of the diffusion process that is based on a stochastic differential equation and delve into an extensive theoretical examination of its implications. Opposed to usual conditional generation tasks, we do not start the reverse process from pure Gaussian noise but from a mixture of noisy speech and Gaussian noise. This matches our forward process which moves from clean speech to noisy speech by including a drift term. We show that this procedure enables using only 30 diffusion steps to generate high-quality clean speech estimates. By adapting the network architecture, we are able to significantly improve the speech enhancement performance, indicating that the network, rather than the formalism, was the main limitation of our original approach. In an extensive cross-dataset evaluation, we show that the improved method can compete with recent discriminative models and achieves better generalization when evaluating on a different corpus than used for training. We complement the results with an instrumental evaluation using real-world noisy recordings and a listening experiment, in which our proposed method is rated best. Examining different sampler configurations for solving the reverse process allows us to balance the performance and computational speed of the proposed method. Moreover, we show that the proposed method is also suitable for dereverberation and thus not limited to additive background noise removal. Code and audio examples are available online[1].**

***Index Terms*—speech enhancement, dereverberation, diffusion models, score-based generative models, score matching.**

I. Introduction

SPEECH enhancement aims to recover clean speech signals from audio recordings that are impacted by acoustic noise or reverberation [1]. To this end, computational approaches often exploit the different statistical properties of the target and interference signals [2]. Machine learning algorithms can be used to extract these statistical properties by learning useful representations from large datasets. A wide class of methods employed for speech enhancement are discriminative models that learn to directly map noisy speech to the corresponding clean speech target [3]. Common approaches include time-frequency (T-F) masking [4], complex spectral mapping [5], or operating directly in the time domain [6]. These supervised methods are trained with a variety of clean/noisy speech pairs containing multiple speakers, different noise types, and a large range of signal-to-noise ratios (SNRs). However, it is nearly impossible to cover all possible acoustic conditions in the training data to guarantee generalization. Furthermore, some discriminative approaches have been shown to result in unpleasant speech distortions that outweigh the benefits of noise reduction [7].

The use of generative models for speech enhancement, on the other hand, follows a different paradigm, namely to learn a prior distribution over clean speech data. Thus, they aim at learning the inherent properties of speech, such as its spectral and temporal structure. This prior knowledge can be used to make inferences about clean speech given noisy or reverberant input signals that are assumed to lie outside the learned distribution. Several approaches follow this idea and utilized deep generative models for speech enhancement [8]–[17]. Among them are methods that employ likelihood-based models for explicit density estimation such as the variational autoencoder (VAE) [18], or leverage generative adversarial networks (GANs) [19] for implicit density estimation. Bando et al. propose a statistical framework using a VAE trained in an unsupervised fashion to learn a prior distribution over clean speech [9]. At test time they combine the speech model with a low-rank noise model to infer the signal variances of speech and noise to build a Wiener filter for denoising. However, since the VAE is trained with clean speech only, the inference model (i.e. the encoder) that predicts the latent variable remains sensitive to noise. This has been shown to cause the generative speech enhancement method to produce speech-like sounds although only noise is present [9]. To mitigate this, it has been proposed to make the inference model robust to noisy speech by training on labeled data in a supervised manner [14], [15], or by disentangling the latent variable from high-level information such as speech activity which can be estimated by supervised classifiers [12], [13]. Nevertheless, VAE-based speech enhancement methods remain limited due to the dimensionality reduction in the latent layer and the combined use of a linear noise model based on non-negative matrix factorization [9]–[15].

More recently, a new class of generative models called diffusion-based generative models, has been introduced to the task of speech enhancement [20]–[23]. Diffusion-based generative models, or simply *diffusion models*, are inspired by non-equilibrium thermodynamics and exist in several variants

This work has been funded by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169), DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the Grant-No. HIDSS-0002, and the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380. We would like to thank J. Berger and Rohde&Schwarz SwissQual AG for their support with POLQA.

Simon Welker is with the Signal Processing Group, Department of Informatics, Universität Hamburg, 22527 Hamburg Germany, and with the Center for Free-Electron Laser Science, DESY, 22607 Hamburg, Germany (e-mail: simon.welker@uni-hamburg.de). The other authors are all with the Signal Processing Group, Department of Informatics, Universität Hamburg, 22527 Hamburg Germany (e-mail: {julius.richter; jeanmarie.lemercier; bunlong.lay; timo.gerkmann}@uni-hamburg.de).

[1]https://github.com/sp-uhh/sgmse

[24]–[26]. All of them share the idea of gradually turning data into noise, and training a neural network that learns to invert this process for different noise scales. More specifically, the inference model is a fixed Markov chain, that slowly transforms the data into a tractable prior, such as the standard normal distribution. The generative model is another Markov chain that is trained to revert this process iteratively [25]. Therefore, diffusion models can be considered as deep latent variable models and have similar properties to VAEs, with the crucial difference that the inference model is not trained and that the latent variables have the same dimensionality as the input. This has the advantage of not relying on surrogate objectives to approximate maximum likelihood training such as the evidence lower bound and enforces no strong restrictions on the model architecture. Recently, diffusion models have been connected with score matching [27] by looking at the stochastic differential equation (SDE) associated with the discrete-time Markov chain [28]. The forward process can be inverted, resulting in a corresponding reverse SDE which depends only on the score function of the perturbed data [29]. Using this continuous-time SDE formalism creates the opportunity to design novel diffusion processes that support the underlying generation task. In contrast to discrete Markov chains, it also allows the use of general-purpose SDE solvers to numerically integrate the reverse process for sampling.

Concerning the application of diffusion models for speech enhancement, there exist currently two approaches that differ conceptually in how the diffusion process is used. One approach is based on speech re-generation, i.e. a diffusion-based vocoder network is used to synthesize clean speech by sampling from an unconditional prior, while a conditioner network takes noisy speech as input and performs the core part of denoising by providing enhanced speech representations to the vocoder network [23], [30]. An auxiliary loss is introduced for the conditioner network to facilitate its ability to estimate clean speech representations [23]. The second approach, on the other hand, does not require any auxiliary loss and is not using two separate models for generation and denoising. Instead, it models the corruption of clean speech by environmental background noise or reverberation directly within the forward diffusion process, so that reversing this process would consequently result in generating clean speech. This has been proposed as a discrete diffusion process for time-domain speech signals [21], and as a continuous SDE-based diffusion process in the complex spectrogram domain [22]. Interestingly, the original denoising score matching objective [31], which is to estimate the white Gaussian noise in the perturbed data, is essentially reminiscent of the goal of speech enhancement, which is to remove interfering noise or reverberation from speech signals. However, under realistic conditions, the environmental noise or reverberation may not match the assumption of stationary white Gaussian noise. Therefore, it was proposed to include real noise recordings in the diffusion process, either by linearly interpolating between clean and noisy speech along the process [21], or by defining such a transformation within the drift term of an SDE [22]. The choice of linear interpolation in [21], however, implies that the trained deep neural network (DNN) must explicitly estimate a portion of environmental noise at each step in the reverse process. This can be seen in the resulting objective function [21, Eq. (21)] which exhibits characteristics of a discriminative learning task. In contrast, an SDE-based formulation results in a pure generative objective function [22, Eq. (9)] and avoids any prior assumptions on the noise distribution.

Nonetheless, note that diffusion-based speech enhancement methods, unlike the VAE-based method described above, are not counted as unsupervised methods, since labeled data (i.e. clean and noisy speech pairs) are used for training. However, the learning objective remains generative in nature which is to learn a prior for clean speech per se rather than a direct mapping from noisy to clean speech. In fact, supervision is only exploited to learn the conditional generation of clean speech when noisy speech is given. Thus, current diffusion-based models for speech enhancement, such as [21]–[23], can be considered as conditional generative models trained in a supervised manner.

In this work, we build upon our previous publication which defines the diffusion process in the complex short-time Fourier transform (STFT) domain [22]. We present a comprehensive theoretical review of the underlying score-based generative model and include an expanded discussion on the conditional generation process which is based on the continuous-time SDE formalism. By using a network architecture developed in the image processing community [28], in the work at hand we significantly improve performance in comparison to our previous model [22]. This indicates that the network, rather than the formalism, was the main limitation of our original approach. In an extensive cross-dataset evaluation, we show that the improved method can compete with recent discriminative models and achieves better generalization when evaluating on a different corpus than used for training. To confirm the effectiveness of the proposed method on non-simulated data, we perform an instrumental evaluation with real-world noisy recordings using non-intrusive metrics. We complement the results with a listening experiment, in which our proposed method is rated best. Interestingly, using the improved network, we show that the proposed method is also suitable for dereverberation when an individual model is trained on simulated reverberant data. Thus, the method is not limited to the removal of additive background noise and can also be applied to non-additive corruptions such as reverberation or, as shown in [32], for bandwidth extension. Furthermore, we investigate different sampler configurations for solving the reverse process which reveals a trade-off between the performance and computational speed of the proposed method.

We summarize our major contributions as follows. Regarding the novelty with respect to Song et al. [28], we introduce a drift term to the SDE to achieve the required task adaptation for reconstruction problems and furthermore apply the diffusion process and score matching objective to a complex data representation. Also note that the approach in [28] is not explicitly trained on reconstruction tasks and the application is different from ours. Regarding the novelty with respect to our previous publication [22], we use an improved network architecture and increase the performance significantly. Moreover, we include

an extended theoretical discussion and investigate different sampler configurations. Finally, we expand the evaluation by means of a cross-dataset evaluation, an instrumental evaluation with real-world noisy recordings, and a listening experiment.

II. Method: Score-based Generative Model for Speech Enhancement (SGMSE)

In this section, we motivate and describe in detail the approach of using score-based generative models for speech enhancement, as proposed in our previous publication [22].

A. Data representation

We represent our data in the complex-valued STFT domain, as it has been observed that both real and imaginary parts of clean speech spectrograms exhibit clear structure and are therefore amenable to deep learning models [4]. Following the approach of complex spectral mapping [5], we use our conditional generative model to estimate the clean real and imaginary spectrograms from the noisy ones.

The use of complex coefficients as data representation allows the definition of the diffusion process in the complex spectral domain, in which additive Gaussian noise corresponds to the signal model used for the denoising task. This relates to traditional STFT-based methods, where spectral coefficients are usually assumed to be complex Gaussian distributed and mutually independent [1], [2]. Statistical approaches often consider an additive signal model assuming that the speech process and the noise process are realizations of stochastic processes that are statistically independent. Observing that the overall noise process is a sum of several independent sources, the central limit theorem ensures that the observed noise process tends to be Gaussian [1].

Although it would be theoretically possible to define the diffusion process in the magnitude domain, additive Gaussian noise would not relate to the signal model anymore. This becomes evident considering that in the magnitude domain, additive Gaussian noise could result in negative amplitudes which are physically not defined.

Thus, we operate on complex spectrograms that are elements of $\mathbb{C}^{T\times F}$, where T denotes the number of time frames dependent on the audio length, and F represents the number of frequency bins. To compensate for the typically heavy-tailed distribution of STFT speech amplitudes [33], we apply an amplitude transformation

$$\tilde{c} = \beta |c|^{\alpha} \mathrm{e}^{i \angle(c)} \tag{1}$$

to all complex STFT coefficients c, where $\angle(\cdot)$ represents the angle of a complex number, $\alpha \in (0, 1]$ is a compression exponent which brings out frequency components with lower energy (e.g. fricative sounds of unvoiced speech) [34], and $\beta \in \mathbb{R}_+$ is a simple scaling factor to normalize amplitudes roughly to within $[0, 1]$. Such a compression has been argued to be perceptually more meaningful in speech enhancement [35], [36], and the transformation ensures that the neural network operates on consistently scaled inputs with respect to the Gaussian diffusion noise [25].

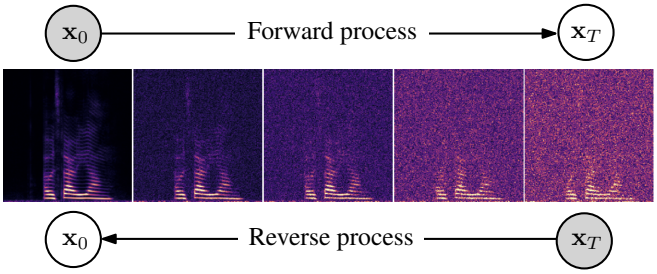


Fig. 1: Diffusion process on a spectrogram: In the forward process noise is gradually added to the clean speech spectrogram $\mathbf{x}_0$, while the reverse process learns to generate clean speech in an iterative fashion starting from the corrupted signal $\mathbf{x}_T$.

B. Stochastic Process

The tasks at hand, speech enhancement and dereverberation, can be considered as conditional generation tasks: Given the corrupted noisy/reverberant speech, generate clean speech by using a conditional generative model. Most previously published diffusion-based generative models are adapted to such conditional tasks either through explicit conditioning channels added to the DNN [37], [38], or through combining an unconditionally trained score model with a separate model (such as a classifier) that provides conditioning in the form of a gradient [28], [39]. With our method, we explore a third possibility, which is to incorporate the particular task directly into the forward and reverse processes of a diffusion-based generative model.

a) Forward Process: Following Song et al. [28], we design a stochastic diffusion process $\{\mathbf{x}_t\}_{t=0}^{T}$ that is modeled as the solution to a linear SDE of the general form,

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, \mathbf{y})\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}\,, \tag{2}$$

where $\mathbf{x}_t$ is the current process state, $t \in [0, T]$ a continuous time-step variable describing the progress of the process (not to be confused with the time index of any signal in the time or T-F domain), $\mathbf{y}$ the noisy or reverberant speech, and $\mathbf{w}$ denotes a standard Wiener process. The vector-valued function $\mathbf{f}(\mathbf{x}_t, \mathbf{y})$ is referred to as the *drift coefficient*, while $g(t)$ is called the *diffusion coefficient* and controls the amount of Gaussian white noise injected at each time-step. Note that different to Song et al. [28], our drift term is now a function of $\mathbf{y}$, by which we tailor the proposed SDE to reconstruction tasks. The process is defined for each T-F bin independently. Thus, the variables in bold are assumed to be vectors in $\mathbb{C}^d$ with $d = TF$ containing the coefficients of a flattened complex spectrogram.

The forward process in Eq. (2) turns a clean speech sample $\mathbf{x}_0$ into a corrupted sample $\mathbf{x}_T$ by gradually adding noise from the Wiener process, as illustrated in Fig. 1. To account for the intended task adaptation of speech enhancement or dereverberation, we propose a drift term that ensures the mean of the process moving from clean speech $\mathbf{x}_0$ to noisy/reverberant speech $\mathbf{y}$. In particular, we define the drift coefficient $\mathbf{f}$ and the diffusion coefficient g as

$$\mathbf{f}(\mathbf{x}_t, \mathbf{y}) := \gamma(\mathbf{y} - \mathbf{x}_t)\,, \tag{3}$$

$$g(t) := \sigma_{\text{min}} \left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\right)^t \sqrt{2 \log \left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\right)}, \tag{4}$$

where γ is a constant called *stiffness* controlling the transition from $\mathbf{x}_0$ to $\mathbf{y}$, and σ_{min} and σ_{max} are parameters defining the noise schedule of the Wiener process. Note that we choose the diffusion coefficient identical to that of the so-called *Variance Exploding SDE* from Song et al. [28]. Our novel contribution lies in the modified drift term, by which the intended task adaptation is achieved.

b) Reverse Process: Following Anderson [29] and Song et al. [28], the SDE in Eq. (2) has an associated *reverse SDE*,

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, \mathbf{y}) + g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})\right] \mathrm{d}t + g(t) \mathrm{d}\bar{\mathbf{w}}\,, \tag{5}$$

where the *score* $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})$ is the term to be approximated by a DNN which is therefore called a *score model*. We denote the score model as $\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t)$, which is parameterized by a set of parameters θ and receives the current process state $\mathbf{x}_t$, the noisy speech $\mathbf{y}$, and the current time-step t as an input. Finally, by substituting the score model into the reverse SDE in Eq. (5), we obtain the so-called *plug-in reverse SDE* [40],

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, \mathbf{y}) + g(t)^2 \mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t)\right] \mathrm{d}t + g(t) \mathrm{d}\bar{\mathbf{w}}\,, \tag{6}$$

which can be solved by various solver procedures, to be discussed in detail in Sec. III.

For inference, we assume that a trained score model $\mathbf{s}_\theta$ is given, which approximates the true score for all $t \in [0, T]$. We can then generate clean speech $\mathbf{x}_0$ conditioned on the noisy or reverberant speech $\mathbf{y}$ by solving the plug-in reverse SDE in Eq. (6). To determine the initial condition of the reverse process at $t = T$, we sample

$$\mathbf{x}_T \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_T; \mathbf{y}, \sigma(T)^2 \mathbf{I}), \tag{7}$$

which is a strongly corrupted version of the noisy speech $\mathbf{y}$, as illustrated in Fig. 1. The denoising process which solves the task of speech enhancement or dereverberation is then based on iterating through the reverse process starting at $t = T$ and ending at $t = 0$.

C. Training objective

Next, we derive the objective function used for training the score model $\mathbf{s}_\theta$. Since the SDE in Eq. (2) describes a Gaussian process, the mean and variance of the process state $\mathbf{x}_t$ can be derived when its initial conditions are known [41]. This allows for direct sampling of $\mathbf{x}_t$ at an arbitrary time step t given $\mathbf{x}_0$ and $\mathbf{y}$ by using the so-called *perturbation kernel*,

$$p_{0t}(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_t; \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t), \sigma(t)^2 \mathbf{I}\right), \tag{8}$$

where $\mathcal{N}_{\mathbb{C}}$ denotes the circularly-symmetric complex normal distribution and $\mathbf{I}$ denotes the identity matrix. We utilize Eqs. (5.50, 5.53) in Särkkä & Solin [41] to determine closed-form solutions for the mean

$$\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t) = \mathrm{e}^{-\gamma t} \mathbf{x}_0 + (1 - \mathrm{e}^{-\gamma t}) \mathbf{y}\,, \tag{9}$$

and the variance

$$\sigma(t)^2 = \frac{\sigma_{\text{min}}^2 \left(\left(\sigma_{\text{max}}/\sigma_{\text{min}}\right)^{2t} - \mathrm{e}^{-2\gamma t}\right) \log(\sigma_{\text{max}}/\sigma_{\text{min}})}{\gamma + \log(\sigma_{\text{max}}/\sigma_{\text{min}})}\,. \tag{10}$$

Vincent [31] shows that fitting the score model $\mathbf{s}_\theta$ to the score of the perturbation kernel $\nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y})$ is equivalent to implicit and explicit score matching [27] under some regularity conditions. This technique is called *denoising score matching* and essentially results in estimating

$$\nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) = \nabla_{\mathbf{x}_t} \log \left[|2\pi\sigma\mathbf{I}|^{-\frac{1}{2}} \mathrm{e}^{-\frac{\|\mathbf{x}_t - \boldsymbol{\mu}\|_2^2}{2\sigma^2}}\right] \tag{11}$$

$$= \nabla_{\mathbf{x}_t} \log |2\pi\sigma(t)\mathbf{I}|^{-\frac{1}{2}} - \nabla_{\mathbf{x}_t} \frac{\|\mathbf{x}_t - \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t)\|_2^2}{2\sigma(t)^2} \tag{12}$$

$$= -\frac{\mathbf{x}_t - \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t)}{\sigma(t)^2}\,, \tag{13}$$

where for simplicity we derived the score for the real and imaginary part of the complex normal distribution in Eq. (8), assuming they are independently distributed and each follows a real-valued multivariate normal distribution. Note that Eq. (13) involves division by $\sigma(t)^2$, which has very small numerical values (including 0) around $t = 0$. To avoid undefined values and numerical instabilities, we thus introduce a small minimum process time t_ε, as done previously in the literature [28].

At each training step, the procedure can then be described as follows: 1) sample a random $t \sim \mathcal{U}[t_\varepsilon, T]$, 2) sample $(\mathbf{x}_0, \mathbf{y})$ from the dataset, 3) sample $\mathbf{z} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{z}; 0, \mathbf{I})$, and 4) sample $\mathbf{x}_t$ from Eq. (8) by effectively computing

$$\mathbf{x}_t = \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t) + \sigma(t)\mathbf{z}. \tag{14}$$

After passing $(\mathbf{x}_t, \mathbf{y}, t)$ to the score model, the final loss is an unweighted L_2 loss between the model output and the score of the perturbation kernel. By substituting Eq. (14) into Eq. (13), the overall training objective becomes,

$$\arg\min_\theta \mathbb{E}_{t, (\mathbf{x}_0, \mathbf{y}), \mathbf{z}, \mathbf{x}_t|(\mathbf{x}_0, \mathbf{y})} \left[\left\| \mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t) + \frac{\mathbf{z}}{\sigma(t)} \right\|_2^2\right], \tag{15}$$

where the expectation is approximated by sampling all random variables at each training step as described above. Note that due to the cancellation of $\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t)$, the loss function does not explicitly involve $\mathbf{y}$, only as an input to the score model. This means that the score model is not tasked with estimating any portion of the environmental noise directly. Finally, the minimization is achieved by optimizing the parameters θ using stochastic gradient descent.

D. Interpretation and limitations

Let p_t be the distribution of the perturbed data $\mathbf{x}_t$ from the diffusion process for a given dataset. Then its time evolution can be thought of as a continuum of distributions $\{p_t\}_{t \in [0,T]}$ which is determined by the drift and the diffusion coefficient of the forward SDE. For fixed $\mathbf{x}_0$ and $\mathbf{y}$, this time evolution can be described in close form using Eq. (8), which we illustrate in Fig. 2 for a one-dimensional case. In the reverse process, the DNN has the task of learning this continuous family of distributions starting from $\tilde{p}_T$ as defined in Eq. (7). Due to the exponential increase of the diffusion coefficient in the forward process with initial condition $\sigma(0)^2 = 0$ the distribution p_0 essentially corresponds to the clean speech distribution, whereas the terminating distribution p_T is a strongly corrupted version

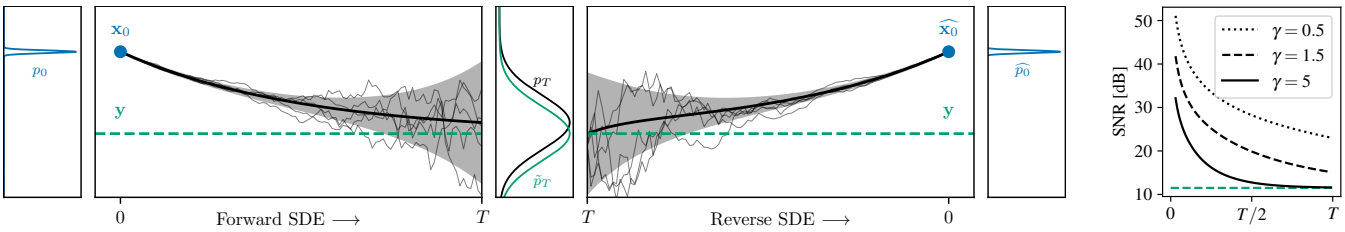


Fig. 2: (Left) The forward and reverse process illustrated with a single scalar variable. The mean $\boldsymbol{\mu}$ (thick black line) of the forward process exponentially decays from clean speech $\mathbf{x}_0$ (blue) towards noisy speech $\mathbf{y}$ (green), and the standard deviation (shaded gray region) increases exponentially. The reverse process moves back to $\mathbf{x}_0$, starting from a slightly mismatched distribution $\tilde{p}_T$ which is centered around $\mathbf{y}$ rather than $\mathbf{x}_T$. Sample paths from both processes are shown as thin black lines. (Right) Time evolution of the SNR of the mean $\boldsymbol{\mu}$ (black) with respect to the SNR of $\mathbf{y}$ (green) for three different values of γ.

of the noisy speech distribution. The particular characteristics in each noisy speech sample are strongly masked by the Gaussian white noise at $t = T$. Therefore, by learning the reverse process, the generative model learns a strong prior p_0 on clean speech, whereas the forward process terminates in a strongly corrupted distribution of the noisy speech, used as a weakly informative prior for generation. In Fig. 2, we simulate five sample paths from the diffusion process. All sample paths of the forward process start exactly at $\mathbf{x}_0$ but exhibit starkly different trajectories at large t. The reverse process should then turn a high-variance sample $\mathbf{x}_T$ back into a low-variance estimate of $\mathbf{x}_0$.

Eq. (9) indicates that the mean $\boldsymbol{\mu}$ of the forward process exponentially decays from $\mathbf{x}_0$ to $\mathbf{y}$, which can also be seen in Fig. 2 (thick black line). However, for finite t it does not fully reach the corrupted speech $\mathbf{y}$ (dashed green line), particularly we have $\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, T) \neq \mathbf{y}$. Thus, the final distribution of the forward process p_T exhibits a slight mismatch to the initial distribution of the reverse process $\tilde{p}_T$. We can make this mismatch arbitrarily small by either choosing a high stiffness parameter γ or by increasing σ_{max} to further smooth the density functions of both distributions. However, increasing γ would bring the mean close to $\mathbf{y}$ within a short time of the forward process, which may lead to an unstable reverse process because only the last steps are concerned with removing environmental noise. This effect can be seen in Fig. 2 (right plot), where we plot the SNR of the process mean $\boldsymbol{\mu}$ averaged over 256 randomly selected files from the dataset for three values of γ. Note that we calculated the SNR in the time domain as the ratio of the power of clean speech to the power of environmental noise after inverting the non-linear amplitude transformation of Eq. (1). We see that while for $\gamma = 5$ the mismatch at $t = T$ becomes virtually zero, the change in SNR occurs mainly in the first half of the process. For $\gamma = 0.5$, on the other hand, the mismatch is already more than 10 dB. However, the slope in the SNR is still apparent at the end of the process. Therefore, there is a trade-off to consider when choosing γ which depends on the dataset to be used. Increasing σ_{max} would come at the cost of more reverse iterations since the more white Gaussian noise is added, the less high-level information about the structure of the speech is preserved to serve as a guide in the reverse process. In the experiments, we choose a set of parameters based on empirical hyperparameter optimization.

III. Numerical SDE Solvers

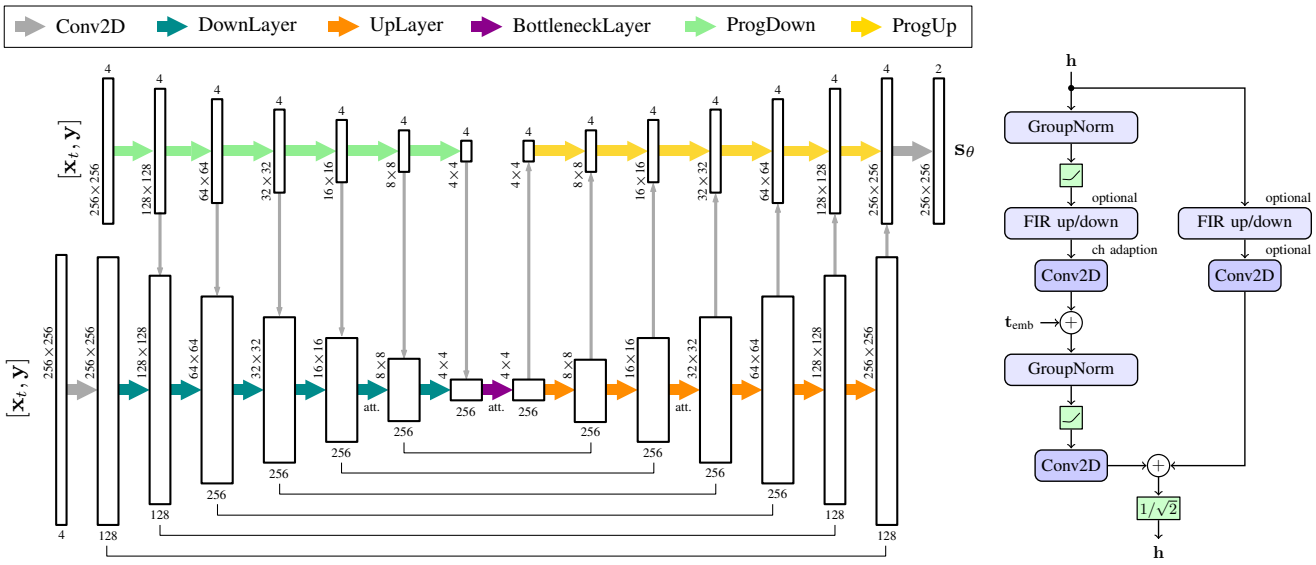
There exist several computational methods to find numerical solutions for SDEs, which are based on an approximation to discrete time steps. To this end, the interval $[0, T]$ is partitioned into N equal subintervals of width $\Delta t = T/N$, which approximates the continuous formulation into the discrete reverse process $\{\mathbf{x}_T, \mathbf{x}_{T-\Delta t}, \ldots, \mathbf{x}_0\}$. A common single-step method for solving this discretization is the Euler-Maruyama method. In each iteration step, the method refers to a previous state of the process and utilizes the drift and the Brownian motion to determine the current state.

In this work, we employ so-called predictor-corrector (PC) samplers proposed by Song et al. [28], which combine single-step methods for solving the reverse SDE with numerical optimization approaches such as annealed Langevin Dynamics [26]. PC samplers consist of two parts, a predictor and a corrector. The predictor can be any single-step method that aims to solve the reverse process by iterating through the reverse SDE. After each iteration step of the predictor, the current state of the process is refined by the corrector. The correction is based on Markov chain Monte Carlo sampling and can be understood as a stochastic gradient ascent optimizer that adds at each iteration step a small amount of noise after taking a step in the direction of the estimated score. One possible intuition about the use of stochastic correctors is that they allow the process state to escape local minima by the use of stochasticity. However, Karras et al. [42] have recently argued that in the reverse process, stochasticity is only necessary to correct for numerical truncation errors of the predictor, a need that could be effectively circumvented by further improving the quality of the score model and predictor.

Another numerical way of approximating the reverse process is by solving the corresponding *probability flow ordinary differential equation (ODE)*,

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, \mathbf{y}) + g(t)^2 \mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t)\right] \mathrm{d}t\,, \qquad (16)$$

which is the associated *deterministic process* of the stochastic reverse SDE in Eq. (6). It can be shown that for each diffusion process, there exists an ODE that describes the same marginal probability density $p_t(\mathbf{x}_t)$ [28]. Enhancing the noisy or reverberant mixture is then based on solving this ODE. In Sec. V-E, we also evaluate and compare this class of solvers



(a) Feature maps at multiple resolutions. (b) Residual block.

Fig. 3: NCSN++ network architecture used as a score model $\mathbf{s}_\theta$: The architecture is based on a multi-resolution U-Net structure containing skip connections and an additional progressive growing path as shown in (a). Each up- and downsampling layer and the bottleneck layer consist of multiple residual blocks in series which are illustrated in (b).

for our task, specifically employing the Runge-Kutta method of fourth order with an error estimator of fifth order [43].

IV. Network Architecture

We utilize the *Noise Conditional Score Network* (NCSN++) architecture [28] for the score model $\mathbf{s}_\theta$ and adapt it for the use of complex spectrograms. For this purpose, we consider the real and imaginary parts of the complex input as separate channels, since the original network only works with real-valued numbers. Estimating both the real and imaginary parts of the score allows to generate complex spectrograms of clean speech.

The network is based on a multi-resolution U-Net structure, which has been experimentally shown to be powerful for tasks such as generation and segmentation [44]. In Fig. 3a, we illustrate the architecture by showing the feature maps at each resolution, indicating their spatial dimension and the corresponding number of channels. The transformations between the feature maps are represented by arrows, where the color of the arrow specifies the type of transformation (see the legend on top). We use Conv2D layers with a 3x3 kernel and stride 1 as input and output layers, and 1x1 Conv2D layers to aggregate information from the progressive growing path that we describe later. Up- and downsampling layers are based on residual network blocks which are taken from the BigGAN architecture [45], shown in Fig. 3b. A residual block consists of Conv2D layers with the same configuration as above, group normalization [46], up- or downsampling with finite impulse response (FIR) filters [47], and the Swish activation function [48]. Each upsampling layer consists of three residual blocks and each downsampling layer of two residual blocks in series with the last block performing the up- or downsampling. Global attention mechanisms [49] are added at a resolution of 16×16 and in the bottleneck layer to better learn global dependencies within the feature maps.

To make the model time-dependent, information about the current progression of the diffusion process is fed into the network architecture. A common practice is to use Fourier-embeddings [49], i.e., a learned projection that maps the scalar time coordinate t to an M-dimensional vector $\mathbf{t}_{\text{emb}}$ that is integrated into every residual block as can be seen in Fig. 3b.

In addition to the main feature extraction path of the multi-resolution U-Net structure, the network incorporates a so-called progressive growing of the input which is seen at the top of Fig. 3a. The idea is to provide a downsampled version of the input to every feature map in the contracting path, which has been successful in stabilizing high-resolution image generation [50]. Note that the downsampling operation in the progressive growing use shared weights for each resolution. The same procedure is also used in the expansive path where a progressive growing of the output is informed by the feature maps at each resolution, resulting in the final score estimate.

V. Experiments

In this section, we describe the experimental setup for our speech enhancement and speech dereverberation experiments using the proposed method.

A. Datasets

For the evaluation of the speech enhancement task we use two datasets, the WSJ0-CHiME3 dataset and the VB-DMD dataset, which are described below. The use of two datasets allows cross-dataset evaluation, i.e. the test is performed on the other dataset than the one used for training. This mismatched condition reveals information about how well the

method generalizes to unseen data with different characteristics such as distinct noise types or different recording conditions. Moreover, to train and evaluate our proposed method on the dereverberation task, we create the WSJ0-REVERB dataset, which is also described below.

a) WSJ0-CHiME3: We create the WSJ0-CHiME3 dataset using clean speech utterances from the Wall Street Journal (WSJ0) dataset [51] and noise signals from the CHiME3 dataset [52]. The mixture signal is created by randomly selecting a noise file and adding it to a clean utterance. Each utterance is used only once, and the SNR is sampled uniformly between 0 and 20 dB for the training, validation, and test set.

b) VB-DMD: We use the publicly available VoiceBank-DEMAND dataset (VB-DMD) [53] which is often used as a benchmark for single-channel speech enhancement. The utterances are artificially contaminated with eight real-recorded noise samples from the DEMAND database [54] and two artificially generated noise samples (babble and speech shaped) at SNRs of 0, 5, 10, and 15 dB. The test utterances are mixed with different noise samples at SNR levels of 2.5, 7.5, 12.5, and 17.5 dB. We split the training data into a training and validation set using speakers "p226" and "p287" for validation.

c) WSJ0-REVERB: To create the WSJ0-REVERB dataset, we use clean speech data from the WSJ0 dataset [51] and convolve each utterance with a simulated room impulse response (RIR). We use the `pyroomacoustics` engine [55] to simulate the RIRs. The reverberant room is modeled by sampling uniformly a T_{60} between 0.4 and 1.0 seconds. A dry version of the room is generated with the same geometric parameters but a fixed absorption coefficient of 0.99, to generate the corresponding anechoic target. The resulting average direct-to-reverberant ratio (DRR) is around -9 dB.

B. Instrumental evaluation metrics

To evaluate the performance of the proposed method we use standard metrics which we will describe in detail below. Metrics (a)-(d) employ full reference algorithms that rate the processed signal in relation to the clean reference signal using conventional digital signal analysis. On the other hand, metrics (e)-(g) are non-intrusive metrics that can be used to evaluate real recordings when the clean reference is unavailable.

a) POLQA: The Perceptual Objective Listening Quality Analysis (POLQA) is an ITU-T standard that includes a perceptual model for predicting speech quality [56]. The POLQA score takes values from 1 (poor) to 5 (excellent) as usual for mean opinion scores (MOS).

b) PESQ: The Perceptual Evaluation of Speech Quality (PESQ) is used for objective speech quality testing and is standardized in ITU-T P.862 [57]. Although it is the predecessor of POLQA, it is still widely used in the research community. The PESQ score lies between 1 (poor) and 4.5 (excellent) and there exist two variants, namely wideband PESQ and narrowband PESQ denoted as $\mathrm{PESQ}_{\mathrm{nb}}$.

c) ESTOI: The Extended Short-Time Objective Intelligibility (ESTOI) is an instrumental measure for predicting the intelligibility of speech subjected to various kinds of degradation [58]. The metric is normalized and lies between 0 and 1, with higher values indicating better intelligibility.

d) SI-SDR, SI-SIR, SI-SAR: Scale-Invariant (SI-) Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR) are standard evaluation metrics for single-channel speech enhancement and speech separation [59]. They are all measured in dB, with higher values indicating better performance.

e) DNSMOS: The Deep Noise Suppression MOS (DNS-MOS) is a reference-free metric to evaluate perceptual speech quality [60]. The evaluation method uses a DNN that is trained on human ratings obtained by using an online framework for listening experiments [61] based on ITU-T P.808 [62].

f) SIG, BAK, OVRL: The non-intrusive speech quality assessment model DNSMOS P.835 [63] is based on a listening experiment according to ITU-T P.835 [64] and provides three MOS scores: speech quality (SIG), background noise quality (BAK), and the overall quality (OVRL) of the audio.

g) WVMOS: Wav-to-Vec MOS (WVMOS) [65] is a MOS prediction method for speech quality evaluation using a fine-tuned wav2vec2.0 model [66].

C. Listening Experiment

Instrumental evaluation metrics do not always correlate to human perception because there are many aspects of perception that are very difficult to capture by computational means. Therefore, we conduct a MUSHRA listening experiment [67] with ten participants using the webMUSHRA framework [68]. The participants were asked to rate the overall quality of twelve randomly sampled examples from the WSJ0-CHiME3 test set as reconstructed by the compared algorithms. The results are reported on a quality scale from 0 to 100.

D. Hyperparameters and training configuration

a) Input representation: We convert each audio input with sampling rate 16 kHz into a complex-valued STFT representation using a window size of 510, resulting in $F = 256$, a hop length of 128 (i.e. approximately 75% overlap), and a periodic Hann window. To process multiple examples for batch training, the length of each spectrogram is trimmed to $T = 256$ STFT time frames, with start and end times selected randomly at each training step. For the spectrogram transformation in Eq. (1), we have chosen $\alpha = 0.5$ and $\beta = 0.15$ empirically.

b) Stochastic process: The SDE in Eq. (2) is parameterized with $\sigma_{\mathrm{min}} = 0.05, \sigma_{\mathrm{max}} = 0.5$, and $\gamma = 1.5$ based on hyperparameter optimization with grid search.

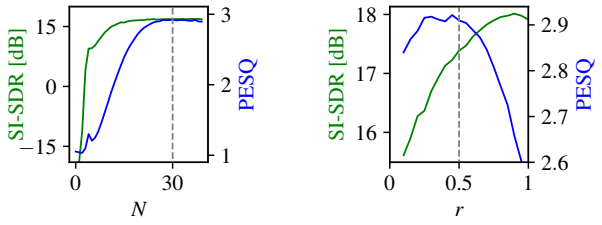
c) Training configuration: We train the DNN on four Quadro RTX 6000 (24 GB memory each) for 160 epochs using the distributed data-parallel (DDP) approach in PyTorch Lightning [69], which takes about one day. We use the Adam optimizer [70] with a learning rate of 10^{-4} and an effective batch size of $4 \times 8 = 32$. We track an exponential moving average of the DNN weights with a decay of 0.999, to be used for sampling [71]. We log the average PESQ value of 20 randomly chosen examples from the validation set during training and select the best-performing model for evaluation.

E. Sampler settings

To find optimal sampler settings for the reverse process, we run a hyperparameter search using the VB-DMD dataset.

TABLE I: Results for different sampler configurations tested on VB-DMD with the average number of function evaluations (NFE) and the respective average real-time factor (RTF)[2].

Type	Sampler settings	NFE	RTF	PESQ	SI-SDR [dB]
PC	0 corrector steps	30	0.89	2.80	15.38
PC	1 corrector steps	60	1.77	2.93	17.35
PC	2 corrector steps	90	2.65	2.92	17.52
ODE	atol=10^{-1}, rtol=10^{-1}	14	0.46	2.78	12.83
ODE	atol=10^{-6}, rtol=10^{-3}	49	1.55	2.71	12.76



(a) Varying N, fix $r = 0.33$ (b) Varying r, fix $N = 30$

Fig. 4: Model performance in PESQ and SI-SDR as a function of (a) the number of reverse steps N and (b) the step size parameter r in the annealed Langevin corrector.

a) Sampler type: We investigate which choice of sampler yields the best speech enhancement performance, comparing the PC sampler with different numbers of corrector steps and an ordinary ODE sampler as described in Sec. III. In Tab. I we can see that use of one correction step in the PC sampler seems to be advantageous, but the use of two steps does not lead to a further increase in performance. Thus, we decide to use the PC sampler with one corrector step for the evaluation. However, it should be noted that the use of one correction step doubles the number of function evaluations (NFE) of the sampler. The *function* being the expensive score model, this results in an average real-time factor (RTF) of 1.77, i.e., 1 sec of audio requires 1.77 sec of processing[2]. Comparing the PC sampler with the ODE sampler, we find that the PC sampler performs better in both metrics. However, with suitable settings, the ODE sampler requires only 14 NFE on average which results in an improved RTF of only 0.46.

b) Number of reverse steps N: The number of reverse steps N can be used to set a balance between the computational effort and the performance of the model. In Fig. 4a, we show the speech enhancement performance as a function of N. It can be seen that SI-SDR starts to stagnate earlier than PESQ. We opt for a value of $N = 30$, at which both metrics show no further increase in performance.

c) Step size in corrector: In Fig. 4b, we vary the step size r of the annealed Langevin dynamics in the corrector. Interestingly, this parameter represents a compromise between PESQ and SI-SDR. We choose $r = 0.5$ to achieve a maximum PESQ value while still obtaining a good value for SI-SDR.

[2]Average processing time for 10 audio files on an NVIDIA GeForce RTX 2080 Ti GPU, in a machine with an Intel Core i7-7800X CPU @ 3.50GHz.

F. Baselines

We compare the performance of our proposed method with four generative and four discriminative baselines which we describe in more detail below. All methods are re-trained by us, except for DVAE, MetricGAN+, and CDiffuSE on VB-DMD, for which we obtained the pre-trained model from the authors who used the exact same training data.

a) STCN [11]: A generative VAE-based speech enhancement method which uses a stochastic temporal convolutional network (STCN) [72] that allows the latent variables to have both hierarchical and temporal dependencies. The parameters of the noise model and the latent variables are estimated using a Monte Carlo expectation maximization (MCEM) algorithm.

b) DVAE [17]: Generative speech enhancement method based on an unsupervised dynamical VAE (DVAE) [73] which models temporal dependencies between successive observable and latent variables. Parameters are updated at test time using a variational expectation maximization (VEM) method where the encoder is fine-tuned using stochastic gradient ascent.

c) CDiffuSE [21]: Most related to our proposed method is CDiffuSE, a generative speech enhancement method based on a conditional diffusion process defined in the time domain.

d) SGMSE [22]: Score-based Generative Model for Speech Enhancement (SGMSE) is our previous publication on which the proposed method is based. The main difference is that it uses a deep complex U-Net [74] instead of the NCSN++ architecture as the score model.

e) MetricGAN+ [75]: A discriminative speech enhancement method that uses a generator network for mask-based prediction of clean speech and introduces a discriminator network trained to approximate the PESQ score.

f) Conv-TasNet [76]: An end-to-end neural network that estimates a mask that is used for filtering a learned representation of the noisy mixture. The filtered representation is transformed back to the time domain by a learned decoder.

g) GaGNet [77]: This neural network is trained on a hybrid complex-domain and magnitude-domain regression objective for single-channel dereverberation. It uses so-called "glance" and "gaze" (GaG) modules, which respectively perform a coarse estimation of the magnitude and refine it with phase estimation in the complex domain.

h) TCN+SA+S [78]: This single-channel dereverberation approach uses a self-attention module to extract features from the input magnitude. This representation is then used by a temporal convolutional network followed by a single-layer convolutional smoother that outputs a magnitude estimate, which is used as the training objective. Griffin-Lim iterations are used to reconstruct the phase.

VI. Results

A. Speech Enhancement

In Tab. II, we report the speech enhancement results on the WSJ0-CHiME3 test set for the matched and mismatched condition, i.e. when the training set was also WSJ0-CHiME3 or when the training set was VB-DMD. We compare our proposed method, which we call *SGMSE+*, with selected baseline methods and sort the results by the type of algorithm,

TABLE II: Speech enhancement results obtained for WSJ0-CHiME3 under matched and mismatched training conditions. Values indicate mean and standard deviation. Methods are sorted by the algorithm type, generative (G) or discriminative (D).

Method	Type	Training set	POLQA	PESQ	ESTOI	SI-SDR [dB]	SI-SIR [dB]	SI-SAR[dB]	DNSMOS
Mixture	-	-	2.64 ± 0.68	2.01 ± 0.55	0.81 ± 0.12	13.5 ± 4.7	18.7 ± 5.5	15.4 ± 4.7	3.34 ± 0.37
STCN [11]	G	WSJ0	2.64 ± 0.68	2.01 ± 0.55	0.81 ± 0.12	13.5 ± 4.7	18.7 ± 5.5	15.4 ± 4.7	3.34 ± 0.37
RVAE [17]	G	WSJ0	2.97 ± 0.63	2.31 ± 0.55	0.85 ± 0.11	15.8 ± 5.0	21.6 ± 6.1	17.6 ± 4.9	3.61 ± 0.29
CDiffuse [21]	G	WSJ0-C3	3.08 ± 0.58	2.27 ± 0.51	0.83 ± 0.09	9.2 ± 2.3	19.8 ± 5.9	10.0 ± 2.3	3.43 ± 0.32
SGMSE [22]	G	WSJ0-C3	2.98 ± 0.60	2.28 ± 0.57	0.86 ± 0.09	14.8 ± 4.3	25.4 ± 5.6	15.3 ± 4.2	3.70 ± 0.27
SGMSE+	G	WSJ0-C3	$\mathbf{3.73 \pm 0.53}$	2.96 ± 0.55	0.92 ± 0.06	18.3 ± 4.4	$\mathbf{31.1 \pm 4.6}$	18.6 ± 4.5	$\mathbf{3.99 \pm 0.19}$
MetricGAN+ [75]	D	WSJ0-C3	3.52 ± 0.61	$\mathbf{3.03 \pm 0.45}$	0.88 ± 0.08	10.5 ± 4.5	24.5 ± 5.1	10.7 ± 4.6	3.67 ± 0.30
Conv-TasNet [76]	D	WSJ0-C3	3.65 ± 0.54	2.99 ± 0.58	$\mathbf{0.93 \pm 0.05}$	$\mathbf{19.9 \pm 4.3}$	29.2 ± 4.6	$\mathbf{20.6 \pm 4.5}$	3.79 ± 0.27
STCN [11]	G	VB	2.53 ± 0.66	1.80 ± 0.45	0.79 ± 0.12	11.9 ± 4.5	17.3 ± 4.9	13.8 ± 4.6	3.40 ± 0.34
RVAE [17]	G	VB	2.84 ± 0.61	2.08 ± 0.49	0.82 ± 0.11	13.9 ± 4.8	19.5 ± 5.9	15.8 ± 4.7	3.52 ± 0.31
CDiffuse [21]	G	VB-DMD	2.15 ± 0.57	1.79 ± 0.42	0.71 ± 0.11	3.2 ± 3.2	21.8 ± 7.0	3.4 ± 3.2	3.17 ± 0.29
SGMSE [22]	G	VB-DMD	2.66 ± 0.58	1.94 ± 0.47	0.81 ± 0.11	13.3 ± 4.3	23.5 ± 6.0	13.8 ± 4.2	3.76 ± 0.25
SGMSE+	G	VB-DMD	$\mathbf{3.43 \pm 0.61}$	$\mathbf{2.48 \pm 0.58}$	$\mathbf{0.90 \pm 0.07}$	$\mathbf{16.2 \pm 4.1}$	$\mathbf{28.9 \pm 4.6}$	$\mathbf{16.4 \pm 4.1}$	$\mathbf{4.00 \pm 0.19}$
MetricGAN+ [75]	D	VB-DMD	2.47 ± 0.67	2.13 ± 0.53	0.76 ± 0.12	6.8 ± 3.1	22.9 ± 4.9	7.0 ± 3.1	3.51 ± 0.29
Conv-TasNet [76]	D	VB-DMD	3.13 ± 0.60	2.40 ± 0.53	0.88 ± 0.08	15.2 ± 3.9	26.5 ± 4.6	15.6 ± 4.0	3.68 ± 0.30

which is either generative or discriminative. Considering the matched condition in the upper half of Tab. II, we see that SGMSE+ outperforms all other generative methods in all metrics. Note that STCN and RVAE are both unsupervised speech enhancement methods, i.e. they are trained on clean speech only (WSJ0 or VB). RVAE shows competitive results for SI-SAR, however, its VEM optimization algorithm is very time-consuming due to the fine-tuning of the encoder at test time, resulting in a RTF of $>$10000. This is significant in contrast to STCN with a RTF of 0.64 and SGMSE+ with a RTF of 1.77[2]. Although both VAE-based methods model temporal dependencies, they are limited in their ability to produce high-quality speech, likely due to the dimensionality reduction of the latent variable and the encoder's sensitivity to noisy input, which causes the latent variable to be incorrectly initialized [15].

Comparing SGMSE+ to our previous model SGMSE, we find a significant improvement, especially for the perceptual metrics. We report improvements of 0.75 for POLQA and 0.68 for PESQ. This shows that the proposed generative diffusion process benefits significantly from the adapted network architecture. In our previous paper [22], we have already shown improvements over CDiffuSE for SGMSE in SI-SDR and SI-SAR, which we now back up with also reporting an improvement in ESTOI and DNSMOS and on par results in PESQ and POLQA. With SGMSE+, these improvements become even more significant, e.g. with 0.65 improvement in POLQA and 9.1 dB in SI-SDR compared to CDiffuSE. In qualitative analysis, we found that SGMSE+ is more accurate than CDiffuSE in preserving the high frequencies of fricatives after the completion of the reverse process. To compensate for that, CDiffuSE combines the enhanced files with the original noisy speech signal at a ratio of 0.2 for the final prediction [21]. This results in a trade-off between noise removal and the conservation of the signal. In our proposed approach, on the other hand, we found no significant suppression of high frequencies after completing the reverse process. Therefore, it is not necessary to mix back the noisy mixture to improve the signal quality, resulting in a significantly higher SI-SIR.

The comparison with Conv-TasNet and MetricGAN+ shows that SGMSE+ can keep up with the performance of discriminative methods and even surpasses them in terms of POLQA, SI-SIR, and DNSMOS. Discriminative methods are based on regression problems that optimize certain point-wise loss functions between the corrupted speech and a clean speech reference. For Conv-TasNet and MetricGAN+ these loss functions correspond to established intrusive metrics, namely SI-SDR for Conv-TasNet and PESQ for MetricGAN+. Note that both these discriminative methods shine in particular on the respective metric they used as a loss function. In contrast, generative methods like SGMSE+ are usually not trained to achieve the exact reconstruction of the reference clean speech but rather aim at generating a realization of speech that is on the manifold of clean speech. Thus, we suggest the use of non-intrusive metrics as a complementary measure since they allow an estimation of speech quality without relying on the exact reconstruction of a reference signal. In fact, for the non-intrusive metric DNSMOS our proposed method yields a significantly higher value than the discriminative baselines, indicating the strong ability of our generative model to generate high-quality clean speech.

Looking at the results for the mismatched condition in the bottom half of Tab. II, a general trend of decreasing metrics can be seen for all methods when compared to the corresponding values of the matched condition. This was to be expected since particular properties of the mismatched test set, such as distinct noise types or different recording characteristics of the clean speech have not been seen during training. However, generative methods generally show less degradation in the mismatched condition than discriminative methods. CDiffuSE is an exception, as this method shows significant degradation in the mismatched case. Informal listening reveals a problem with gain control, which is evident in strong volume fluctuations in the enhanced files. Furthermore, we see that SGMSE+ outperforms all other methods in all metrics under this condition, which shows the ability of our proposed method to generalize well.

Complementary to the average results above, we present in

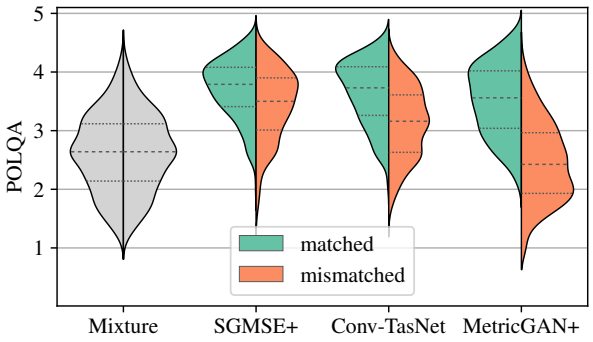


Fig. 5: Violin plots showing POLQA results for the matched and the mismatched condition with dashed and dotted lines representing median and quartiles, respectively.

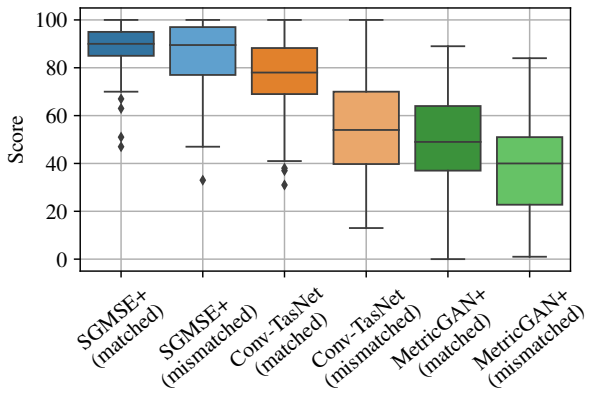


Fig. 6: Boxplot showing the results of the MUSHRA listening experiment with ten participants on twelve randomly selected examples.

Fig. 5 violin plots of the full distribution of the POLQA scores obtained for SGMSE+, Conv-TasNet, MetricGAN+, and the noisy mixture for reference. For each method, the distributions are plotted side by side for the matched and mismatched conditions, so that the ability to generalize can be inferred from the horizontal alignment between both distributions. It can be seen that both distributions for SGMSE+ are relatively similar, whereas they are skewed for Conv-TasNet and especially for MetricGAN+.

In Fig. 6, we report the results of the MUSHRA listening experiment in a boxplot. On average, the ten participants rated the overall quality of our proposed approach with the highest score. In addition, our method remains fairly robust when the model was trained on a different training set, while discriminative methods show much stronger degradation for the mismatched condition. This also corresponds with the results of the non-intrusive metric DNSMOS in Tab. II and thus supports the use of non-intrusive methods for instrumental evaluation. Interestingly, MetricGAN+ was only rated with a median score less than 50 for the matched condition, although the method performed best among all baselines for PESQ (see Tab. II). This reveals the discrepancy between the use of instrumental metrics for evaluation and people's actual perceptions. We suspect that MetricGAN+ has simply learned to utilize the internal operations of the PESQ algorithm to obtain a high value in this metric, neglecting the naturalness of the clean speech estimate. In fact, listening to the enhanced files, it can be recognized that the energy of the speech signal estimated by MetricGAN+ is mainly concentrated in the low- and mid-frequency area of the spectrogram, while high frequencies are strongly attenuated.

Listening to the enhanced files of our method, we notice that at very low input SNRs, some "vocalizing" artifacts with very poor articulation and no linguistic meaning are occasionally produced. In other examples, we find that breathing sounds or speech-like sounds were generated in noisy regions where no speech was originally present. These artifacts may also explain the outliers of our method in the listening experiment (see Fig. 6). For the matched condition, for example, the two lowest outliers come from the same utterance with clearly noticeable vocalizing artifacts. We hypothesize that these artifacts can be linked to the generative nature of the proposed approach.

TABLE III: Speech enhancement results obtained for VB-DMD. Models marked with an asterisk (*) are additional baselines with values taken from the corresponding papers.

Method	PESQ	PESQ_{nb}	ESTOI	SI-SDR	DNSMOS
Mixture	1.97	2.88	0.79	8.4	3.09
SEGAN* [8]	2.16	-	-	-	-
RVAE [17]	2.43	3.11	0.81	16.4	3.30
MetricGAN-U* [82]	2.45	-	0.77	8.2	-
CDiffuse [21]	2.52	3.31	0.79	12.4	3.09
SGMSE [22]	2.28	3.22	0.80	16.2	3.46
SGMSE+	2.93	**3.66**	**0.87**	17.3	**3.56**
UMX* [83]	2.35	-	0.83	14.0	-
Conv-TasNet [76]	2.63	3.42	0.85	**19.1**	3.37
MetricGAN+ [75]	**3.13**	3.63	0.83	8.5	3.37

Indeed, for very noisy inputs, the score model may erroneously identify noise energy in some T-F areas as corrupted speech. The reverse diffusion process then produces speech where it did not originally exist. We argue that this behavior could be mitigated if some conditioning with respect to speech activity and phoneme identity would be added to the score model.

Finally, Tab. III lists the results for the standardized VB-DMD dataset. This has the advantage that one can take values from other methods and copy them from the corresponding papers for a quick algorithmic comparison. It can be seen that SGMSE+ outperforms all other generative baselines, further narrowing the performance gap with discriminative methods that currently lead the benchmark based on PESQ[3], including recent approaches such as [79] and [80]. It should however be noted that PESQ formally requires a minimum file length of 3.2 sec according to P.862.3 [81], which is not the case for most files in VB-DMD [53].

Investigating whether phase estimation has actually been improved with the modeling of the complex coefficients, we use the noisy phase in place of the estimated phase which does not show a significant performance difference. This is also in line with a recent study on the role of phase enhancement, where it has been shown that the impact of phase enhancement is rather small for ∼32 ms spectral analysis frames but increasingly

[3]https://paperswithcode.com/sota/speech-enhancement-on-demand

TABLE IV: Single-channel dereverberation results obtained for WSJ0-REVERB test set. Values indicate mean and standard deviation. Methods are sorted by the algorithm type which is either generative (G) or discriminative (D).

Method	Type	POLQA	PESQ	ESTOI	SI-SDR	SI-SIR	SI-SAR
Mixture	-	1.76 ± 0.29	1.36 ± 0.19	0.46 ± 0.12	-7.3 ± 5.5	-7.5 ± 5.4	-
SGMSE [22]	G	1.79 ± 0.28	1.35 ± 0.15	0.57 ± 0.07	-7.4 ± 5.8	-1.1 ± 7.0	-6.2 ± 5.5
SGMSE+	G	**3.24 ± 0.46**	**2.66 ± 0.48**	**0.84 ± 0.07**	**1.6 ± 7.8**	9.4 ± 10.2	**2.3 ± 7.2**
Conv-TasNet [76]	D	2.41 ± 0.52	1.84 ± 0.42	0.73 ± 0.10	**1.6 ± 5.3**	**12.1 ± 5.1**	1.9 ± 5.4
TCN+SA+S [78]	D	2.92 ± 0.33	2.29 ± 0.36	0.79 ± 0.05	-4.4 ± 5.3	-2.3 ± 5.2	-0.6 ± 5.1
GaGNet [77]	D	2.62 ± 0.47	1.98 ± 0.46	0.73 ± 0.08	-0.6 ± 4.9	6.1 ± 3.9	0.4 ± 5.1

large with shorter frame lengths [84].

B. Dereverberation

We report in Tab. IV the performance of our approach when trained and tested on a single-channel dereverberation task. We compare with SGMSE [22] and three discriminative baselines, namely Conv-TasNet [76], GaGNet [77] and TCN+SA+S [78].

Our proposed SGMSE+ approach performs particularly well in terms of instrumental metrics compared to all other baseline models. The low average input DRR of -9 dB constitutes a real challenge for discriminative approaches, which do not manage to separate the reverberation from the target without distorting the target signal, resulting in low-quality scores. On the other hand, our approach benefits from generative modeling and is able to reconstruct speech with very high quality in most cases. When comparing our previous SGMSE model [22] with SGMSE+, we see that for speech dereverberation, the method benefits greatly from the improved network architecture. This effect is even more significant than for additive background noise removal in the speech enhancement task.

In particular, using the proposed approach SGMSE+ on a single-channel dereverberation task does not produce any of the vocalized artifacts observed in the speech enhancement experiments for low input SNRs. Although the reverberant signal is formally decorrelated in the time domain from the target by the randomness of reflections across the room, it still originates from the dry speech source. Therefore, we conjecture that the score model effectively detects whether the energy in a particular time-frequency area is associated with the clean speech nearby that needs to be reconstructed.

C. Evaluation on real data

Complementing the experiments using simulated data, we evaluate the speech enhancement performance on real-world noisy recordings. For real-world noisy recordings, there exists no clean speech reference. Thus, we can only non-intrusive metrics to evaluate the perceptual speech quality which we describe in Sec. V-B (e)-(g). For the evaluation, we use 300 files from the test set of the Deep Noise Suppression (DNS) Challenge 2020 [85]. In Tab. V, we report the results for models that were trained on VB-DMD. It turns out that our proposed method performs better than all other methods in all non-intrusive metrics, demonstrating its robustness to real-world noisy examples. Interestingly, a trend of degradation in speech quality (SIG) can be observed for the discriminative methods, whereas all generative models improve this metric with respect to the mixture. For the background noise quality (BAK) metric, on the other hand, discriminative models seem to perform well, yet our proposed method performs superior. It is important to note that non-intrusive metrics do not require a corresponding clean reference signal and only assess speech quality based on the method's estimate. We hypothesize that our generative model works well on these metrics, as it was trained to generate clean speech. However, "vocalizing" artifacts as mentioned above or phonetic confusions may not be captured with these metrics.

TABLE V: Speech enhancement results obtained for real-world noisy recordings from the DNS Challenge 2020 test set.

Method	DNSMOS	SIG	BAK	OVRL	WVMOS
Mixture	3.05	3.05	2.51	2.26	1.12
RVAE [17]	3.29	3.16	2.91	2.44	1.87
CDiffuse [21]	3.14	3.15	3.19	2.55	1.86
SGMSE [22]	3.38	3.22	3.02	2.52	1.80
SGMSE+	**3.64**	**3.42**	**3.82**	**3.04**	**2.54**
Conv-TasNet [76]	3.07	2.87	3.59	2.52	2.07
MetricGAN+ [75]	3.26	2.88	3.39	2.45	1.52

We provide on our project page[4] some listening examples for all evaluated tasks. Furthermore, we include real reverberant examples from the MC-WSJ-AV dataset [86].

VII. Conclusions

In this work, we built upon our existing work [22] that uses a novel stochastic diffusion process to design a generative model for speech enhancement in the complex STFT domain. We presented an extended theoretical analysis of the underlying score-based generative model and derived in detail the objective function used for training. In further explorations, we considered the time evolution of the conditional diffusion process which revealed a slight mismatch between the forward and reverse process, which can be adjusted with a careful parameterization of the forward SDE.

By using an adopted network architecture, we were able to significantly improve the performance compared to our previous model. In addition, we trained and evaluated the proposed method on the task of speech dereverberation and show significantly superior performance compared to discriminative baseline methods. Hence, we showed that with our proposed

[4]https://uhh.de/inf-sp-sgmse

method, a single framework can be used to train individual models for different distortion types. For the task of speech enhancement, we evaluated performance under matched and mismatched conditions, i.e. when the training and test data were taken from the same or different corpora. For the matched condition, the proposed generative speech enhancement method performs on par with competetive discriminative methods. For the mismatched condition, our method shows strong generalization capabilities and outperforms all baselines in all metrics, as confirmed by a listening experiment. In very adverse conditions, however, we observe that the proposed method sometimes introduces vocalizing and breathing artifacts. We argue that these could be mitigated in future work if some conditioning concerning speech activity and phoneme information would be added to the score model.

In addition, we explored different sampling strategies to solve the reverse process at test time which allows us to balance the performance and computational speed of the proposed method. Future work could include other sampling techniques to further reduce the number of diffusion steps [87] and thus the computational complexity.

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, 2013.

[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. John Wiley & Sons, 2018.

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

[4] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 24, no. 3, pp. 483–492, 2015.

[5] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," *IEEE Int. Workshop on Machine Learning for Signal Proc. (MLSP)*, pp. 1–6, 2017.

[6] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," *IEEE Asia-Pacific Signal and Inf. Proc. Assoc. Annual Summit and Conf. (APSIPA ASC)*, 2017.

[7] P. Wang, K. Tan, and D. L. Wang, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 28, pp. 39–48, 2020.

[8] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," *ISCA Interspeech*, pp. 3642–3646, 2017.

[9] 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," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 716–720, 2018.

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

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

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

[13] ——, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement," *IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics (WASPAA)*, pp. 126–130, 2021.

[14] Y. Bando, K. Sekiguchi, and K. Yoshii, "Adaptive neural speech enhancement with a denoising variational autoencoder." *ISCA Interspeech*, pp. 2437–2441, 2020.

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

[16] A. A. Nugraha, K. Sekiguchi, and K. Yoshii, "A flow-based deep latent variable model for speech spectrogram modeling and enhancement," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 28, pp. 1104–1117, 2020.

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

[18] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *Int. Conf. on Learning Representations (ICLR)*, 2014.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 27, 2014.

[20] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," *IEEE Asia-Pacific Signal and Inf. Proc. Assoc. Annual Summit and Conf. (APSIPA ASC)*, pp. 659–666, 2021.

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

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

[23] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv preprint arXiv:2206.03065*, 2022.

[24] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *Int. Conf. on Machine Learning (ICML)*, pp. 2256–2265, 2015.

[25] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[26] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 32, 2019.

[27] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching." *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[28] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *Int. Conf. on Learning Representations (ICLR)*, 2021.

[29] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[30] Y. Koizumi, H. Zen, K. Yatabe, N. Chen, and M. Bacchiani, "SpecGrad: Diffusion probabilistic model based neural vocoder with adaptive noise spectral shaping," *ISCA Interspeech*, pp. 803–807, 2022.

[31] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[32] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[33] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," *Int. Workshop on Acoustic Echo and Noise Control*, 2010.

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

[35] C. H. You, S. N. Koh, and S. Rahardja, "/spl beta/-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 13, no. 4, pp. 475–486, 2005.

[36] C. Breithaupt and R. Martin, "Analysis of the decision-directed snr estimator for speech enhancement with respect to low-snr and transient conditions," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 19, no. 2, pp. 277–289, 2010.

[37] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," *Int. Conf. on Learning Representations (ICLR)*, 2021.

[38] G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann, "Conditional image generation with score-based diffusion models," *arXiv preprint arXiv:2111.13606*, 2021.

[39] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 34, 2021.

[40] C.-W. Huang, J. H. Lim, and A. C. Courville, "A variational perspective on diffusion-based generative models and score matching," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 34, 2021.

[41] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.

[42] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 35, 2022.

[43] J. R. Dormand and P. J. Prince, "A family of embedded Runge-Kutta formulae," *Journal of Computational and Applied Mathematics*, vol. 6, pp. 19–26, 1980.

[44] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Int. Conf. on Medical image computing and computer-assisted intervention*, pp. 234–241, 2015.

[45] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *Int. Conf. on Learning Representations (ICLR)*, 2018.

[46] Y. Wu and K. He, "Group normalization," *Proc. of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

[47] R. Zhang, "Making convolutional networks shift-invariant again," *Int. Conf. on Machine Learning (ICML)*, pp. 7324–7334, 2019.

[48] P. Ramachandran, B. Zoph, and Q. V. Le, "Swish: a self-gated activation function," *arXiv preprint arXiv:1710.05941*, 2017.

[49] 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 Inf. Proc. Systems (NeurIPS)*, vol. 30, 2017.

[50] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 8110–8119, 2020.

[51] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete." [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S6A

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

[53] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," *ISCA Speech Synthesis Workshop (SSW)*, pp. 146–152, 2016.

[54] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[55] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2018.

[56] ITU-T Rec. P.863, "Perceptual objective listening quality prediction," *Int. Telecom. Union (ITU)*, 2018. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[57] 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," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, vol. 2, pp. 749–752, 2001.

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

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

[60] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 6493–6497, 2021.

[61] B. Naderi and R. Cutler, "An open source implementation of ITU-T recommendation P.808 with validation," *ISCA Interspeech*, pp. 2862–2866, 2020.

[62] ITU-T Rec. P.808, "Subjective evaluation of speech quality with a crowdsourcing approach," *Int. Telecom. Union (ITU)*, 2021. [Online]. Available: https://www.itu.int/rec/T-REC-P.808-202106-I/en

[63] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2022.

[64] ITU-T Rec. P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *Int. Telecom. Union (ITU)*, 2003. [Online]. Available: https://www.itu.int/rec/T-REC-P.835-200311-I/en

[65] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: a unified framework for bandwidth extension and speech enhancement," *arXiv preprint arXiv:2203.13086*, 2022.

[66] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, 2020.

[67] ITU-R Rec. BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems," *Int. Telecom. Union (ITU)*, 2014. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1534

[68] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webmushra—a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.

[69] W. Falcon et al., "Pytorch lightning," *GitHub: https://github.com/PyTorchLightning/pytorch-lightning*, vol. 3, 2019.

[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int. Conf. on Learning Representations (ICLR)*, 2015.

[71] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, pp. 12 438–12 448, 2020.

[72] E. Aksan and O. Hilliges, "Stcn: Stochastic temporal convolutional networks," *Int. Conf. on Learning Representations (ICLR)*, 2018.

[73] 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.

[74] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware Speech Enhancement with Deep Complex U-Net," *arXiv preprint arXiv:1903.03107*, 2019.

[75] 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," *arXiv preprint arXiv:2104.03538*, 2021.

[76] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 27, no. 8, pp. 1256–1266, 2019.

[77] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, vol. 187, p. 108499, 2022.

[78] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 28, pp. 1598–1607, 2020.

[79] G. Yu, A. Li, H. Wang, Y. Wang, Y. Ke, and C. Zheng, "DBT-Net: Dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 30, pp. 2629–2644, 2022.

[80] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based metric GAN for speech enhancement," in *ISCA Interspeech*, 2022, pp. 936–940.

[81] ITU-T Rec. P.862.3, "Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2," *Int. Telecom. Union (ITU)*, 2007. [Online]. Available: https://www.itu.int/rec/T-REC-P.862.3/en

[82] 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," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 7412–7416, 2022.

[83] S. Uhlich and Y. Mitsufuji, "Open-unmix for speech enhancement (UMX SE)," 2020. [Online]. Available: https://github.com/sigsep/open-unmix-pytorch

[84] T. Peer and T. Gerkmann, "Phase-aware deep speech enhancement: It's all about the frame length," *JASA Express Letters*, vol. 2, no. 10, p. 104802, 2022.

[85] 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," *ISCA Interspeech*, pp. 2492–2496, 2020.

[86] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 357–362, 2005.
[87] D. Watson, W. Chan, J. Ho, and M. Norouzi, "Learning fast samplers for diffusion models by differentiating through sample quality," *Int. Conf. on Learning Representations (ICLR)*, 2021.

A.4 Wind Noise Reduction with a Diffusion-based Stochastic Regeneration Model [P9]

Abstract

In this paper we present a method for single-channel wind noise reduction using our previously proposed diffusion-based stochastic regeneration model combining predictive and generative modelling. We introduce a non-additive speech in noise model to account for the non-linear deformation of the membrane caused by the wind flow and possible clipping. We show that our stochastic regeneration model outperforms other neural-network-based wind noise reduction methods as well as purely predictive and generative models, on a dataset using simulated and real-recorded wind noise. We further show that the proposed method generalizes well by testing on an unseen dataset with real-recorded wind noise. Audio samples, data generation scripts and code for the proposed methods can be found online.

Reference

Jean-Marie Lemercier, Joachim Thiemann, Raphael Koning and Timo Gerkmann "Wind Noise Reduction with a Diffusion-based Stochastic Regeneration Model", *VDE 15th ITG Conference on Speech Communication*, 2023, DOI: 10.30420/456164022

Copyright Notice

Authors' Contributions

Jean-Marie Lemercier is the first author of this publication. He implemented all algorithms, trained the neural networks used in the paper, conducted the experimental validation, and wrote the manuscript. Timo Gerkmann brought insights on the experimental validation, mathematical derivations, reviewed the manuscript and presented the poster at VDE 15th ITG Conference on Speech Communication. Joachim Thiemann and Raphael Koning brought their valuable feedback.

Wind Noise Reduction with a Diffusion-based Stochastic Regeneration Model

Jean-Marie Lemercier,⋆ Joachim Thiemann,† Raphael Koning,† Timo Gerkmann⋆

⋆ {first.name}@uni-hamburg.de, Universität Hamburg, Hamburg, Germany
† {first.name}@advancedbionics.com, Advanced Bionics, Hannover, Germany

Abstract

In this paper we present a method for single-channel wind noise reduction using our previously proposed diffusion-based stochastic regeneration model combining predictive and generative modelling. We introduce a non-additive speech in noise model to account for the non-linear deformation of the membrane caused by the wind flow and possible clipping. We show that our stochastic regeneration model outperforms other neural-network-based wind noise reduction methods as well as purely predictive and generative models, on a dataset using simulated and real-recorded wind noise. We further show that the proposed method generalizes well by testing on an unseen dataset with real-recorded wind noise. Audio samples, data generation scripts and code for the proposed methods can be found online[1][2].

1 Introduction

Wind noise captured in microphone signals is an important factor of intelligibility and quality loss in speech communications, and occurs for virtually all outdoor scenarios. Hearing-device users particularly suffer from wind noise presence, more than from other noise types [1]. Wind acoustics are highly non-stationary, especially in case of strong wind as it adopts a turbulent behaviour close to microphones. Furthermore, the corruption caused by wind noise exhibit non-linear behaviours, due to the displacement of the microphone membrane by the air flow and saturation for high wind noise levels [2]. Such non-stationarity and non-linearities make enhancing speech corrupted with wind noise a very difficult challenge [2, 3].

Several traditional enhancement solutions leverage multi-channel processing [4] and often exploit the spatial coherence structure across microphones shaped by the local turbulent flow [2, 5, 6]. Single-channel solutions include adaptive post-filtering [7] and spectral enhancement exploiting the particular spectrum of wind noise [8]. Other approaches were designed using the fact that wind noise resides mostly in low-frequency regions. These methods discard the polluted low-frequency speech information, and aim to recreate a clean version of it based on artificial bandwidth extension or synthesis techniques [9, 10].

More recently, machine learning solutions were proposed [11, 12], mostly relying on supervised predictive learning, i.e. recovering clean speech from noisy speech based on a mapping learnt by a deep neural network (DNN) during training. Generative models are a different class of machine learning techniques that learn a parameterization of the clean speech distribution and allow to generate multiple valid estimates instead of a single best estimate as for predictive approaches [13]. Such generative methods include variational auto-encoders (VAEs), normalizing flows, generative adversarial networks (GANs) and diffusion models [14]. Diffusion models were recently proposed for speech restoration tasks such as enhancement, dereverberation and bandwidth extension [15–18]. Originally intended for image generation, they showed impressive results on speech restoration, notably outperforming their predictive counterparts on speech quality [18]. In previous work [19], we proposed to combine predictive and generative modelling to leverage both the fast inference and interference removal power of predictive approaches, and the sample quality and generalization abilities of generative models. The resulting model was evaluated on additive noise and dereverberation separately.

We aim here to investigate the performance of the proposed model for wind noise reduction. We introduce a signal model approximation for speech in wind noise taking into account possible non-linearities such as membrane displacement and clipping which often occur for strong winds [2]. We show that our stochastic regeneration model is able to highly increase the quality and intelligibility of speech in wind noise. We compare to DNN-based baselines for wind noise reduction, as well as purely generative and predictive models using the same DNN architecture as the proposed method. We validate our algorithm on both the matched test split of our simulated dataset and an unseen speech in wind noise dataset using real-recorded wind noise samples.

2 Diffusion-based generative models

Diffusion models are a class of generative models that iteratively generate data from noise based on a stochastic process parameterization [14, 20]. More specifically, they use a forward diffusion process during training to progressively degrade clean data with Gaussian noise and/or other types of corruption. At inference time, a reversed version of the diffusion process generates a sample from the target data distribution given an initial Gaussian noise state.

2.1 Forward and reverse processes

The stochastic forward process $\{\mathbf{x}_\tau\}_{\tau=0}^{T}$ is defined as a stochastic differential equation (SDE) [20]:

$$\mathrm{d}\mathbf{x}_\tau = \mathbf{f}(\mathbf{x}_\tau, \tau)\mathrm{d}\tau + g(\tau)\mathrm{d}\mathbf{w}, \tag{1}$$

where $\mathbf{x}_\tau$ is the current state of the process indexed by the continuous time step $\tau \in [0, T]$. This *diffusion time* variable τ relates to the progress of the stochastic process and should not be mistaken for our usual notion of time in time-series-like signals. The initial condition represents target clean speech $\mathbf{x}_0 = \mathbf{x}$. As our process is defined in the complex spectrogram domain, independently for each time-frequency (T-F) bin, the variables in bold are assumed to be vectors in $\mathbb{C}^d$ containing the coefficients of the complex spectrogram— with d the product of the time and frequency dimensions— whereas variables in regular font

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

[2] This work has been funded by the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380. The authors are responsible for the content of this paper.

represent real scalar values. The stochastic process $\mathbf{w}$ is a standard d-dimensional Brownian motion, that is, $\mathrm{d}\mathbf{w}$ is a zero-mean Gaussian random variable with standard deviation $\mathrm{d}\tau$ for each T-F bin. The *drift* function $\mathbf{f}$ and *diffusion* coefficient g as well as the initial condition $\mathbf{x}_0$ and the final diffusion time T uniquely define the process $\{\mathbf{x}_\tau\}_{\tau=0}^{T}$ Under some regularity conditions on $\mathbf{f}$ and g, the reverse process $\{\mathbf{x}_\tau\}_{\tau=T}^{0}$ is another diffusion process and is also the solution of a SDE [20, 21]:

$$\mathrm{d}\mathbf{x}_\tau = \left[-\mathbf{f}(\mathbf{x}_\tau,\tau) + g(\tau)^2 \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)\right] \mathrm{d}\tau + g(\tau)\mathrm{d}\bar{\mathbf{w}}, \quad (2)$$

where $\mathrm{d}\bar{\mathbf{w}}$ is a d-dimensional Brownian motion for the time flowing in reverse and $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$ is the *score function*, i.e. the gradient of the logarithmic data distribution for the current process state $\mathbf{x}_\tau$.

In order to perform speech restoration , the generation of clean speech $\mathbf{x}$ is conditioned on cues depending on the noisy speech $\mathbf{y}$. Previous diffusion-based approaches proposed to condition the process explicitly within the neural network [22] or through guided classification [23]. In [16] however, it has been proposed to include the conditioning information directly into the diffusion process by defining the forward process as the solution to the following Ornstein-Uhlenbeck SDE:

$$\mathrm{d}\mathbf{x}_\tau = \underbrace{\gamma(\mathbf{y}-\mathbf{x}_\tau)}_{:=\mathbf{f}(\mathbf{x}_\tau,\mathbf{y})}\mathrm{d}\tau + \underbrace{\left[\sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\tau}\sqrt{2\log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}\right]}_{:=g(\tau)}\mathrm{d}\mathbf{w}. \quad (3)$$

The stiffness hyperparameter γ controls the slope of the decay from $\mathbf{y}$ to $\mathbf{x}_0$, and the noise extrema $\sigma_{\min}$ and $\sigma_{\max}$ control the noise scheduling, i.e. the amount of white Gaussian noise injected at each timestep during the forward process. Therefore, the forward process in Eq. (3), injects an infinitesimal amount of corruption $\gamma(\mathbf{y}-\mathbf{x}_t)\mathrm{d}\tau$ to the current process state $\mathbf{x}_\tau$, along with Gaussian noise with standard deviation $g(\tau)\mathrm{d}\tau$. It is shown in [16] that the solution to (3) admits a complex Gaussian perturbation kernel $p(\mathbf{x}_\tau|\mathbf{x}_0,\mathbf{y})$ with mean $\boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},\tau)$ and variance $\sigma(\tau)^2$:

$$\boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},\tau) = \mathrm{e}^{-\gamma\tau}\mathbf{x}_0 + (1-\mathrm{e}^{-\gamma\tau})\mathbf{y}, \quad (4)$$

$$\sigma(\tau)^2 = \frac{\sigma_{\min}^2\left(\left(\sigma_{\max}/\sigma_{\min}\right)^{2\tau} - \mathrm{e}^{-2\gamma\tau}\right)\log\left(\sigma_{\max}/\sigma_{\min}\right)}{\gamma + \log\left(\sigma_{\max}/\sigma_{\min}\right)}. \quad (5)$$

2.2 Score function estimator

During inference, the score function $\nabla_{\mathbf{x}_\tau}\log p(\mathbf{x}_\tau)$ is not known and must be estimated by a so-called *score model* $\mathbf{s}_\theta$. Once obtained, all quantities are available for solving Eq. (2) with classical numerical methods (see Section 2.3). Given the Gaussian form of the perturbation kernel $p(\mathbf{x}_\tau|\mathbf{x}_0,\mathbf{y})$, the following *denoising score matching* objective can be used to train the score model $\mathbf{s}_\phi$ [24]:

$$\mathscr{J}^{(\mathrm{DSM})}(\phi) = \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{y}),\mathbf{z},\mathbf{x}_\tau}\left[\left\|\mathbf{s}_\phi(\mathbf{x}_\tau,\mathbf{y},\tau) + \tfrac{\mathbf{z}}{\sigma(\tau)}\right\|_2^2\right]. \quad (6)$$

To optimize (6), a clean utterance $\mathbf{x}_0$ and noisy utterance $\mathbf{y}$ are first picked in the training set. A diffusion time step τ is sampled uniformly in $[\tau_\epsilon, T]$ where $\tau_\epsilon > 0$ is a minimal diffusion time used to avoid numerical instabilities. Then the current process state is obtained by Gaussianity of the perturbation kernel as $\mathbf{x}_\tau = \boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},\tau) + \sigma(\tau)\mathbf{z}$, with $\mathbf{z} \sim \mathscr{N}_{\mathbb{C}}(\mathbf{z};\mathbf{0},\mathbf{I})$. Classical gradient descent methods are then used to tune the score model (see Section 4.2).

2.3 Inference through reverse sampling

At inference time, we sample $\mathbf{x}_T$, with:

$$\mathbf{x}_T \sim \mathscr{N}_{\mathbb{C}}(\mathbf{x}_T;\mathbf{y},\sigma^2(T)\mathbf{I}). \quad (7)$$

Conditional generation is then performed by solving the reverse SDE (2) from $\tau = T$ to $\tau = 0$, where the score function is replaced by its estimator $\mathbf{s}_\phi$. We use classical SDE numerical solvers [20] based on a discretization of (2) according to a uniform grid of N points on the interval $[0,T]$ (no minimal diffusion time is needed here). We will denote by G_ϕ the generative model corresponding to reverse diffusion such that the clean speech estimate is $\widehat{\mathbf{x}} = \mathbf{x}_0 = G_\phi(\mathbf{y})$.

3 Stochastic regeneration model

We now revisit our **Sto**chastic **R**egeneration **M**odel (StoRM) combining predictive and generative modelling originally proposed in [19]. An initial predictor D_θ is used as a first stage to generate a denoised version of the sample (see Figure 1). This estimate can be polluted by residual noise and speech distortions due to the the fact that predictive models trained with a mean-square error objective map noisy speech to the posterior mean $\mathbb{E}[\mathbf{x}|\mathbf{y}]$ rather than to a sample of the posterior distribution [13, 19]. A generative diffusion model G_ϕ then learns to regenerate the clean speech $\mathbf{x}_0$ given $D_\theta(\mathbf{y})$:

$$\hat{\mathbf{x}} = G_\phi(D_\theta(\mathbf{y})). \quad (8)$$

The inference process is shown in Figure 1. For training, we use a criterion $\mathscr{J}^{(\mathrm{StoRM})}$ combining denoising score matching $\mathscr{J}^{(\mathrm{DSMS})}$ (where the difference with (6) is the presence of $D_\theta(\mathbf{y})$ as extra-conditioning) and a supervised regularization term $\mathscr{J}^{(\mathrm{Sup})}$ matching the output of the initial predictor to the target speech:

$$\begin{aligned}
\mathscr{J}^{(\mathrm{DSMS})}(\theta) &= \mathbb{E}_{\tau,(\mathbf{x},\mathbf{y}),\mathbf{z}}\left\|\mathbf{s}_\phi(\mathbf{x}_\tau,[\mathbf{y},D_\theta(\mathbf{y})],\tau) + \tfrac{\mathbf{z}}{\sigma(\tau)}\right\|_2^2,\\
\mathscr{J}^{(\mathrm{Sup})}(\phi) &= \mathbb{E}_{(\mathbf{x},\mathbf{y})}\|\mathbf{x} - D_\theta(\mathbf{y})\|_2^2,\\
\mathscr{J}^{(\mathrm{StoRM})}(\theta,\phi) &= \mathscr{J}^{(\mathrm{DSMS})}(\theta) + \alpha\,\mathscr{J}^{(\mathrm{Sup})}(\phi),
\end{aligned} \quad (9)$$

where a value of $\alpha = 1$ is empirically chosen. As $D_\theta(\mathbf{y})$ may not be a sufficient cue for optimal reconstruction of the target speech, we additionally provide $\mathbf{y}$ as conditioning to the score model $\mathbf{s}_\phi$ by stacking it with $D_\theta(\mathbf{y})$ (see Section 4.2).

4 Experimental Setup

4.1 Data

We generate our simulated dataset using clean speech data from the WSJ0 corpus and simulated and recorded wind noise, each making up for half of the noise data. The simulated half of the noise dataset is created with the wind noise generator [25]. Wind noise with airflow speed-dependent behaviour is generated using randomized airflow profiles (see Table 1). The real-recorded other half of the noise dataset is obtained from public sources such as Freesounds

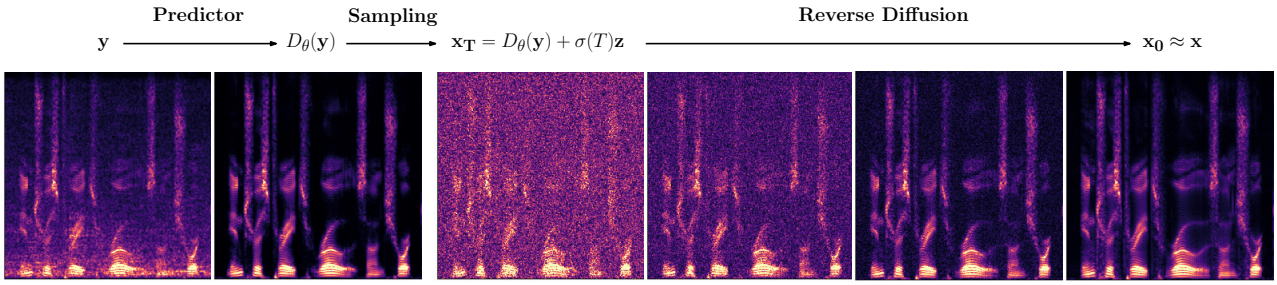


Figure 1: StoRM inference process. The predictive stage produces a denoised version $D_\theta(\mathbf{y})$. Reverse diffusion G_ϕ is then carried out by first adding Gaussian noise $\sigma(T)\mathbf{z}$ to obtain the start sample $\mathbf{x}_T$, and finally by solving the reverse diffusion SDE (2) to obtain the estimated clean speech $\mathbf{x}_0$.

Parameter	Unit	Distribution
Number of wind gusts		$\mathscr{U}(1,10)$
Input SNR	dB	$\mathscr{U}(-6,14)$
Compressor ratio		$\mathscr{U}(1,20)$
Compressor sidechain input level		$\mathscr{U}(0.8,1.2)$
Compressor attack	ms	$\mathscr{U}(5,100)$
Compressor release	ms	$\mathscr{U}(5,500)$
Clipping presence		$\mathscr{B}(0.75)$
Clipping threshold η		$\mathscr{U}(0.85,1.0)$

Table 1: Data generation parameters

(4.3 h), YouTube (0.1 h) and various open-source noise databases (1.8 h) [26–28].

We design a non-additive speech in noise model by taking into account both non-linearities caused by microphone membrane displacement and clipping in case of strong wind. First, wind noise and speech signals are mixed additively with a random SNR. Then, the membrane displacement non-linearity is simulated by using a compressor on the speech signal, sidechained by the noise signal. If the wind noise signal exceeds the compressor threshold, the speech signal is compressed by an amount determined by the compressor ratio and the magnitude of the noise signal above the compressor threshold. We sample compressor threshold, ratio, attack and release parameters to mimic various recording devices. Finally, hard-clipping is simulated by limiting the dynamic range of the noisy signal $\mathbf{y}$ between $-\eta \max(|\mathbf{y}|)$ and $\eta \max(|\mathbf{y}|)$. We refer the reader to Table 1 for the data generation parameters. In total 25, 2.3 and 1.5 hours of noisy speech sampled at 16kHz are created for training, validation and testing respectively. We make our data generation method publicly available[3].

Finally, we also use an unseen dataset using real wind noise recorded in a wind tunnel, added to German speech with a SNR in $\{0, -5, -10\}$ dB. For this data, provided by Advanced Bionics, only noisy speech without ground truth is available.

4.2 Hyperparameters and training setting

Data representation

Noisy and clean utterances are transformed using a short-time Fourier transform (STFT) with a window size of 510, a hop length of 128 and a square-root Hann window, at a sampling rate of 16kHz, as in [17, 19]. A square-root magnitude warping is used to reduce the dynamical range of spectrograms [17]. During training, sequences of 256 STFT frames (≈2s) are extracted from the full-length utterances with random offsets and normalized by the maximum absolute value of the noisy utterance.

Forward and reverse diffusion

For the proposed stochastic regeneration model, we fix the stiffness to γ =1.5, the extremal noise levels to $\sigma_{\text{min}} = 0.05$ and $\sigma_{\text{max}} = 0.5$, and the extremal diffusion times to $T = 1$ and $\tau_\epsilon = 0.03$ as in [19]. $N = 20$ time steps are used for reverse diffusion using the first-order Euler-Maruyama prediction scheme, resulting in 21 neural network calls.

Network architecture

For score estimation and initial prediction, we use two copies of a lighter configuration of the NCSN++ architecture [20], which was proposed in our previous study [18] and denoted as *NCSN++M* and has roughly 27.8M parameters. For initial prediction, the noisy speech spectrogram $\mathbf{y}$ real and imaginary channels are stacked and provided as sole input to the network D_θ, and no noise-conditioning is used. For score estimation during reverse diffusion, the noisy speech spectrogram $\mathbf{y}$, the initial prediction $D_\theta(\mathbf{y})$ and the current estimate $\mathbf{x}_\tau$ real and imaginary channels are stacked and fed to the network $\mathbf{s}_\phi$, and the current noise level $\sigma(\tau)$ is provided as a conditioner. The resulting approach is denoted as *StoRM*.

We also investigate using GaGNet for initial prediction [29], a state-of-the-art predictive denoising approach conducting parallel magnitude- and complex-domain enhancement in the T-F domain. We use 257 frequency bins instead of the original 161 for compatibility with NCSN++-based score estimation, increasing the network capacity to 11.6M parameters compared to the original 5.9M. The resulting approach is denoted as *StoRM-G*.

Baselines

We compare our approaches to the purely generative SGMSE+M [17] and purely predictive NCSN++M [18]. SGMSE+M uses the NCSN++M architecture for score estimation, $N = 30$ reverse time steps with a Euler-Maruyama predictor and one step of Annealed Langevin Dynamics correction with step size $r = 0.5$, resulting in 60 neural network calls. We change the stiffness to $\gamma = 2.5$ and maximal noise level to $\sigma_{\text{max}} = 0.75$. We noticed that higher maximal noise level and stiffness were needed, as the initial mean $\mathbf{y}$ which needs masking by the Gaussian noise $\sigma(T)\mathbf{z}$ has higher energy compared to StoRM where the initial mean is $D_\theta(\mathbf{y})$.

[3] https://github.com/sp-uhh/storm

Method	#Params	DNSMOS	WVMOS	PESQ	ESTOI	SI-SDR
Noisy		3.04 ± 0.61	1.24 ± 2.58	1.70 ± 0.61	0.76 ± 0.19	4.1 ± 5.9
FCN+SANM [11]	4.3 M	2.63 ± 0.66	2.17 ± 1.73	2.01 ± 0.57	0.78 ± 0.15	9.0 ± 4.3
DBLSTM-U [12]	73.5 M	3.50 ± 0.72	3.61 ± 0.49	2.94 ± 0.78	0.90 ± 0.10	15.5 ± 6.5
NCSN++M [18]	27.8 M	4.09 ± 0.39	3.70 ± 0.53	2.76 ± 0.92	$\mathbf{0.92 \pm 0.08}$	$\mathbf{18.8 \pm 6.2}$
SGMSE+M [17]	27.8 M	4.01 ± 0.32	3.79 ± 0.40	2.83 ± 0.78	0.90 ± 0.10	16.5 ± 6.1
StoRM (prop.)	56.0 M	$\mathbf{4.19 \pm 0.30}$	3.80 ± 0.43	3.02 ± 0.76	0.91 ± 0.08	17.4 ± 6.0
StoRM-G (prop.)	39.6 M	$\mathbf{4.19 \pm 0.30}$	$\mathbf{3.87 \pm 0.41}$	$\mathbf{3.07 \pm 0.76}$	$\mathbf{0.92 \pm 0.08}$	17.6 ± 6.0

Table 2: Enhancement results on our simulated test set. Values indicate mean and standard deviation.

Method	DNSMOS	WVMOS
Noisy	1.89 ± 0.41	0.08 ± 0.19
FCN+SANM [11]	1.29 ± 0.33	0.23 ± 0.34
DBLSTM-U [12]	1.96 ± 0.47	0.23 ± 0.33
NCSN++M [18]	3.34 ± 0.59	1.59 ± 0.55
SGMSE+M [17]	3.44 ± 0.11	1.52 ± 0.50
StoRM (prop.)	3.36 ± 0.44	1.33 ± 0.60
StoRM-G (prop.)	$\mathbf{3.56 \pm 0.42}$	$\mathbf{1.67 \pm 0.57}$

Table 3: Enhancement results on the unseen dataset using real-recorded wind noise. Values indicate mean and standard deviation.

We also report the performance of the soft audio noise masking model using fully connected networks (*FCN+SANM*) [11] and the "Unified" version of the deep bidirectional long-short term memory network approach (*DBLSTM-U*) by [12], which is the state-of-the-art DNN-based method for wind noise reduction.

Training configuration

We train the approaches NCSN++M, SGMSE+M, StoRM and StoRM-G using the Adam optimizer [30] with a learning rate of 0.0005 and an effective batch size of 16. We track an exponential moving average of the DNN weights with a decay of 0.999 [31]. We train DNNs for a maximum of 500 epochs using early stopping based on the validation loss with a patience of 10 epochs. For StoRM approaches, the initial predictor is pre-trained with a complex spectrogram mean-square error loss, then we jointly train the predictor and score network with (9) [19]. We implement FCN+SANM and DBLSTM-U using the hyperparameters and training configuration proposed by the authors.

4.3 Evaluation metrics

For instrumental evaluation of the speech enhancement and dereverberation performance with clean test data available, we use intrusive measures such as Perceptual Evaluation of Speech Quality (PESQ) [32] to assess speech quality, Extended Short-Term Objective Intelligibility (ESTOI) [33] for intelligibility and scale-invariant signal to distortion ratio (SI-SDR) [34] for wind noise and distortion removal. For reference-free assessment of speech restoration, we also use the non-intrusive DNSMOS [35] and WVMOS [36] metrics, which perform DNN-based mean opinion score estimation.

5 Results and Discussion

5.1 Simulated dataset

We report in Table 2 instrumental metrics for the proposed method and baselines on the proposed simulated test set. We observe that the FCN+SANM baseline [11] hardly improves over noisy speech, as it uses a simplistic low-capacity architecture without any sequence-modelling module. In comparison, DBLSTM-U [12] yields good results for a simple predictive approach but has a large number of parameters. As already reported in [18, 19], predictive NCSN++M yields high ESTOI and SI-SDR but mediocre quality-related metrics, due to important speech distortions. Purely generative SGMSE+M achieves marginally higher PESQ and WVMOS but lower ESTOI and SI-SDR, and produces many generative artifacts.

The proposed methods StoRM and StoRM-G highly improve speech quality, while remaining competitive with NCSN++M in terms of ESTOI and SI-SDR and using three times fewer operations than SGMSE+M. StoRM-G slightly outperforms StoRM with fewer parameters, showing the efficiency of using GaGNet as initial predictor.

5.2 Real-recorded dataset

We display in Table 3 instrumental metrics of the different baselines and proposed models on the unseen dataset using real-recorded wind noise. We show that NCSN++M generalizes well to unseen noisy data for a predictive approach, compared to the other predicitve baselines. However, SGMSE+M and StoRM-G perform much better, the latter improving DNSMOS by 1.8 points.

6 Conclusions and Future Work

We propose to solve the wind noise reduction task with our previously proposed diffusion-based stochastic regeneration model, combining predictive and generative modelling. We design a speech in noise signal model which deviates from the classical additive model by introducing non-linearities to simulate membrane displacement and clipping. We show that the introduced method is able to strongly increase the quality and intelligibility of speech in wind noise. The proposed stochastic regeneration model outperforms previous DNN-based methods for wind noise reduction as well as purely predictive and generative methods in terms of instrumental metrics. In particular, it generalizes well to unseen data using real-recorded wind noise.

References

[1] S. Kochkin, "Marketrak VIII: Consumer satisfaction with hearing aids is slowly increasing.," *The Hearing Journal*, vol. 63, pp. 19–32, 2010.

[2] C. M. Nelke, *Wind Noise Reduction: Signal Processing Concepts*. PhD thesis, IKS RWTH Aachen, 2016.

[3] J. A. Zakis, "Wind noise at microphones within and across hearing aids at wind speeds below and above microphone saturation," *The Journal of the Acoustical Society of America*, vol. 129, pp. 3897–3907, 06 2011.

[4] P. Thuene and G. Enzner, "Maximum-likelihood approach to adaptive multichannel-wiener postfiltering for wind-noise reduction," in *ITG Symp. Speech Comm.*, Oct. 2016.

[5] D. Mirabilii and E. Habets, "Spatial coherence-aware multichannel wind noise reduction," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 28, pp. 1974–1987, 2020.

[6] S. Franz and J. Bitzer, "Multi-channel algorithms for wind noise reduction and signal compensation in binaural hearing aids," in *Int. Workshop on Acoustic Signal Enhancement*, Aug. 2010.

[7] E. Nemer and W. Leblanc, "Single-microphone wind noise reduction by adaptive postfiltering," in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, Oct. 2009.

[8] C. M. Nelke, N. Chatlani, C. Beaugeant, and P. Vary, "Single microphone wind noise PSD estimation using signal centroids," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2014.

[9] C. M. Nelke, N. Nawroth, M. Jeub, C. Beaugeant, and P. Vary, "Single microphone wind noise reduction using techniques of artificial bandwidth extension," in *Proc. Euro. Signal Proc. Conf. (EUSIPCO)*, Aug. 2012.

[10] C. M. Nelke, P. A. Naylor, and P. Vary, "Corpus based reconstruction of speech degraded by wind noise," in *Proc. Euro. Signal Proc. Conf. (EUSIPCO)*, Mar. 2015.

[11] H. Bai, F. Ge, and Y. Yan, "DNN-based speech enhancement using soft audible noise masking for wind noise reduction," *China Communications*, vol. 15, no. 9, pp. 235–243, 2018.

[12] J. Lee, K. Kim, T. Z. Shabestary, and H.-G. Kang, "Deep bidirectional long short-term memory based speech enhancement for wind noise reduction," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, Mar. 2017.

[13] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[14] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Int. Conf. Machine Learning (ICML)*, Apr. 2015.

[15] 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)*, June 2022.

[16] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Interspeech*, Sept. 2022.

[17] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *arXiv 2208.05830*, 2022.

[18] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing discriminative versus diffusion generative models for speech restoration tasks," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, June 2023.

[19] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *arXiv 2212.11851*, 2022.

[20] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Int. Conf. Learning Repr. (ICLR)*, May 2021.

[21] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[22] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," *Int. Conf. Learning Repr. (ICLR)*, May 2021.

[23] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Neural Information Proc. Systems (NIPS)*, vol. 34, Dec. 2021.

[24] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[25] D. Mirabilii, A. Lodermeyer, F. Czwielong, S. Becker, and E. A. Habets, "Simulating wind noise with airflow speed-dependent characteristics," in *Int. Workshop on Acoustic Signal Enhancement*, Sept. 2022.

[26] IKS RWTH Aachen University, "The IKS wind noise database," 2023. https://www.iks.rwth-aachen.de/forschung/tools-downloads/databases/wind-noise-database.

[27] Yang, "Wind noise dataset," 2023. https://doi.org/10.5281/zenodo.6687982.

[28] K. Arendt, A. Szumaczuk, B. Jasik, K. Piaskowski, P. Masztalski, M. Matuszewski, K. Nowicki, and P. Zborowski, "Test dataset for separation of speech, traffic sounds, wind noise, and general sounds," 2020. https://doi.org/10.5281/zenodo.4279220.

[29] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, vol. 187, p. 108499, 2022.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int. Conf. Learning Repr. (ICLR)*, May 2015.

[31] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *Neural Information Proc. Systems (NIPS)*, Dec. 2020.

[32] 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.

[33] J. Jensen and C. 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.

[34] J. L. 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.

[35] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *arXiv 2010.15258*, 2021.

[36] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "HiFi++: a unified framework for bandwidth extension and speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, June 2023.

A.5 Unsupervised Blind Joint Dereverberation and Room Acoustics Estimation with Diffusion Models [P11]

Abstract

In this paper, we present an unsupervised method for single-channel blind dereverberation and room impulse response estimation. Our algorithm is rooted in Bayesian posterior sampling: it combines a likelihood model enforcing fidelity to the reverberant measurement, and an anechoic speech prior implemented by an unconditional diffusion model. We design a parametric filter to represent the room impulse response, with exponential decays for each frequency subband. Room acoustics estimation and speech dereverberation are jointly carried out, as the filter parameters are iteratively estimated and the speech utterance refined along the reverse diffusion trajectory. In a purely blind scenario where the room impulse response is unknown, we are able to successfully perform speech dereverberation in various acoustic scenarios, significantly outperforming other blind unsupervised baselines. Unlike supervised methods, which often struggle to generalize, our method seamlessly adapts to different acoustic conditions. This paper extends our previous conference work by offering comprehensive experiments and new insights into the algorithm's performance and flexibility. Notably, we demonstrate the adaptability of our method to high-resolution singing voice dereverberation, study the performance of our method on RIR estimation, and conduct subjective evaluation experiments to validate the perceptual quality of the results, among other contributions. Audio samples and code can be found online uhh.de/sp-inf-buddy

Reference

Jean-Marie Lemercier, Eloi Moliner, Simon Welker, Vesa Välimäki and Timo Gerkmann, "Unsupervised Blind Joint Dereverberation and Room Acoustics Estimation with Diffusion Models", *Preprint. Submitted to: IEEE/ACM Trans. Audio, Speech, Language Proc.*, 2024, DOI: 10.48550/arXiv.2408.07472

Copyright Notice

Authors' Contributions

The two first authors Jean-Marie Lemercier and Eloi Moliner have equal contribution. They co-wrote Sections III and IV. Jean-Marie Lemercier conducted the speech dereverberation and robustness experiments, he wrote Sections I, II and V.A. Eloi Moliner conducted the singing voice dereverberation and experiments, the subjective evaluations and the RIR estimation validation, he wrote Sections V.B and V.C. Simon Welker brought some feedback through discussions on the methods developed in the paper. Timo Gerkmann and Vesa Välimäki brought insights on the experimental validation, mathematical derivations, and reviewed the manuscript.

Unsupervised Blind Joint Dereverberation and Room Acoustics Estimation with Diffusion Models

Jean-Marie Lemercier*, *Student Member, IEEE*, Eloi Moliner*, Simon Welker, *Student Member, IEEE*, Vesa Välimäki, *Fellow, IEEE*, Timo Gerkmann, *Senior Member, IEEE*

***Abstract*—This paper presents an unsupervised method for single-channel blind dereverberation and room impulse response (RIR) estimation, called BUDDy. The algorithm is rooted in Bayesian posterior sampling: it combines a likelihood model enforcing fidelity to the reverberant measurement, and an anechoic speech prior implemented by an unconditional diffusion model. We design a parametric filter representing the RIR, with exponential decay for each frequency subband. Room acoustics estimation and speech dereverberation are jointly carried out, as the filter parameters are iteratively estimated and the speech utterance refined along the reverse diffusion trajectory. In a blind scenario where the room impulse response is unknown, BUDDy successfully performs speech dereverberation in various acoustic scenarios, significantly outperforming other blind unsupervised baselines. Unlike supervised methods, which often struggle to generalize, BUDDy seamlessly adapts to different acoustic conditions. This paper extends our previous work by offering new experimental results and insights into the algorithm's performance and versatility. We first investigate the robustness of informed dereverberation methods to RIR estimation errors, to motivate the joint acoustic estimation and dereverberation paradigm. Then, we demonstrate the adaptability of our method to high-resolution singing voice dereverberation, study its performance in RIR estimation, and conduct subjective evaluation experiments to validate the perceptual quality of the results, among other contributions. Audio samples and code can be found online.[1]**

***Index Terms*—Acoustics, diffusion models, reverberation, speech enhancement.**

I. Introduction

REVERBERATION is a natural phenomenon caused by acoustic waves propagating in a room and getting reflected by walls. Reverberation and particularly late reflections often degrade speech intelligibility and quality for normal listeners, and even more severely so for hearing-impaired listeners [1]. Therefore, many communication devices now include a dereverberation algorithm, which aims to recover the anechoic component of speech. This paper considers the scenario in which recordings from only one microphone are available, which is more challenging than a multi-channel scenario [2].

Traditional dereverberation algorithms operate in the time, spectral, or cepstral domain [3], leveraging statistical assumptions about the anechoic and reverberant signals [4] as well as properties of the reverberation signal model [5]. Two scenarios are considered for dereverberation, depending on the knowledge of the room acoustics represented by the room impulse response (RIR). Some methods tackle *informed* scenarios, where the RIR is known [5], [6], whereas other approaches consider *blind* scenarios where the RIR is unknown [7]–[11]. Informed dereverberation is naturally an easier task than blind dereverberation. However, knowing the RIR does not guarantee obtaining a stable and causal inverse filter in the single-channel case, since real-world RIRs are mixed-phase systems [12]. Using multiple microphones helps resolve this issue to some extent [2], but informed dereverberation methods generally exhibit other weaknesses such as a lack of robustness to RIR estimation errors [13]. Additionally, most scenarios in real-life applications are (at least partially) blind, as the RIR is either not measured beforehand, or only valid for a specific acoustic setting.

Data-driven approaches rely less on distributional assumptions than statistical methods but instead directly learn the signal properties and structures from data [14]. Among these, supervised predictive models are particularly popular for blind dereverberation: these range from time-frequency masking [15] and mapping [16] to algorithms operating on the cepstrum [17] or directly on the waveform [18], [19].

Generative modeling is another paradigm gaining a lot of interest in audio restoration tasks [20], including dereverberation. Generative models for speech dereverberation learn a parameterization of the posterior distribution of clean speech conditioned on reverberant speech. Diffusion models in particular [21]–[23] have been extensively investigated for such conditional generation task, leading to the introduction of diffusion-based blind supervised dereverberation algorithms [24], [25]. Still, the generalization ability of supervised approaches is limited by their design.

In contrast, unsupervised methods have been getting little visibility but boast interesting properties such as improved robustness to unseen acoustic conditions without the need for retraining. An unsupervised method for informed single-channel dereverberation based on diffusion models was proposed in our prior work [26]. That approach is based on Bayesian diffusion posterior sampling (DPS) [27], combining a diffusion-based anechoic speech prior and a Gaussian likelihood model for state-of-the-art informed dereverberation. However, as shown in this work, such an informed algorithm is sensitive to even small RIR estimation errors, rendering it impractical in real-life scenarios.

*Equal contribution. Jean-Marie Lemercier, Simon Welker and Timo Gerkmann are with the Signal Processing group at Universität Hamburg, Hamburg, Germany. Eloi Moliner and Vesa Välimäki are with the Acoustics Lab, Department of Information Communications Engineering, Aalto University, Espoo, Finland. The authors gratefully acknowledge the computing resources provided by both the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (NHR project F101AC1) and the Aalto Science-IT project.

[1]uhh.de/sp-inf-buddy

Related works in other signal processing domains have already considered blind inverse problems through the lens of posterior sampling with diffusion priors. For image deblurring, Chung et al. [28] propose to use an additional diffusion process dedicated to estimating the deblurring kernel, while Laroche et al. [29] adapts an expectation-maximization algorithm using a denoising regularization of the blurring kernel, and Sanghvi et al. [30] dedicates a non-blind solver to estimate a deblurred image at each diffusion step. For speech denoising, Nortier et al. [31] combine a noise model based on non-negative matrix factorization with a clean speech diffusion prior. Moliner et al. [32] address the problem of blind bandwidth extension by leveraging a diffusion prior and iteratively optimizing a parametric lowpass filter operator. Recent works adapt denoising diffusion restoration models (DDRM) [33] for singing voice dereverberation [34], [35], using an initialization provided by the weighted-prediction error (WPE) algorithm [7].

For speech dereverberation, a first generative model based on traditional Gaussian mixtures was proposed in [36]. Other works learn an anechoic speech prior via variational auto-encoding (VAE): the VAE-NMF method [37] models reverberation via non-negative matrix factorization and estimates its parameters with a Monte-Carlo method; the RVAE-EM model [38] adopts a maximum a posteriori perspective, combining a recurrent VAE prior with a Gaussian likelihood model. Unsupervised dereverberation with a non-generative prior has also been investigated in the multi-channel scenario [39].

This paper expands our prior work [40], where we designed a blind unsupervised dereverberation algorithm, extending [26] to the blind scenario. The resulting approach, called BUDDy, uses a model-based parametric subband filter with an exponential decay to approximate the RIR. BUDDy performs joint estimation of the RIR and the anechoic speech, leveraging the model-based parameterization as an acoustic prior and the diffusion model as a speech prior. We have shown previously [40] that BUDDy can successfully remove reverberation, and that it is robust to changes in acoustic conditions because of the lack of supervision during training. Therefore, BUDDy closes the performance gap between matched and mismatched acoustic conditions in comparison to diffusion-based supervised approaches [24], [25].

In this paper, we extend the experimental framework of our previous publication [40] with the following contributions:

- In Section II, we investigate the *robustness* of informed dereverberation approaches in partially blind scenarios. We highlight the limitations of these approaches when the RIR is perturbed with Gaussian noise or estimated blindly using a state-of-the-art RIR estimator [41].
- Section V-A extends the evaluation of BUDDy for speech dereverberation beyond instrumental metrics, including a *subjective listening test* and a set of *ablation studies*.
- Section V-B presents new experiments on applying BUDDy to *singing voice dereverberation* at a sampling rate of 44.1 kHz, which is higher compared to the 16-kHz sampling rate used in our speech experiments [40]. The results, which also include a subjective listening test, indicate that our method significantly outperforms existing unsupervised state-of-the-art approaches and performs comparably to a supervised adaptation of the proposed method.
- Finally, Section V-C assesses BUDDy's performance in *RIR estimation* against a state-of-the-art supervised estimator [41]. We use frequency-wise acoustic descriptors to evaluate the accuracy of BUDDy on reverberation time and clarity.

We organize the paper as follows. Section II reports on the robustness of RIR-informed methods, providing context for the proposed approach. In Section III, we introduce diffusion-based generative models and posterior sampling methods for dereverberation using diffusion priors as proposed in previous work [26]. Then in Section IV, we introduce our blind unsupervised dereverberation method BUDDy [40]. The experiments and results mentioned above are presented in Section V. Lastly, Section VI concludes the paper.

II. Robustness of RIR-Informed Methods in Partially Blind Scenarios

We consider dereverberation under the prism of inverse problem solving: we wish to retrieve the anechoic utterance waveform $\mathbf{x}_0 \in \mathbb{R}^L$, where L is the speech utterance length, given the reverberant measurement $\mathbf{y}$. Reverberation is often modelled as a convolution between the anechoic speech with a RIR $\mathbf{h} \in \mathbb{R}^{L_\mathbf{h}}$, such that $\mathbf{y} = \mathbf{h} * \mathbf{x}_0$, where $*$ is the discrete convolution operator in the time domain, resulting in $\mathbf{y} \in \mathbb{R}^{L+L_\mathbf{h}-1}$.

Informed dereverberation algorithms such as [5], [26] assume complete knowledge of the room acoustics as provided by the RIR $\mathbf{h}$. However, as pointed out in Section I, even if the RIR is perfectly known, single channel dereverberation is not trivial as RIRs represent mixed-phase systems such that causal and stable inverse filters do not exist [12]. In practical applications, the RIR is often unknown, and even when it can be estimated, there are typically estimation errors making the task of robust single channel dereverberation even harder. Before delving into blind dereverberation, which is the main focus of this paper, we first examine and demonstrate the sensitivity of informed dereverberation methods in such *partially blind* scenarios, i.e., when RIRs are known up to estimation errors.

We investigate here two methods, the first of which is our prior work InfDerevDPS [26], which is based on Bayesian diffusion posterior sampling (DPS) [27]. InfDerevDPS combines a diffusion-based anechoic speech prior and a Gaussian likelihood model which measures the adherence of the current estimate to the reverberant utterance, given the reverberation operator. The diffusion-based speech prior and sampling technique for InfDerevDPS are presented in Section III-B hereafter. The second method RIF+Post [5] performs regularized inverse filtering in the Fourier domain, followed by traditional speech enhancement.

We start studying the case where the oracle RIR is corrupted by Gaussian noise. The results displayed in Fig. 1 indicate that the performance of both the diffusion-based and the traditional method dwindles as the noise power increases. This suggests that informed methods have very limited robustness to errors

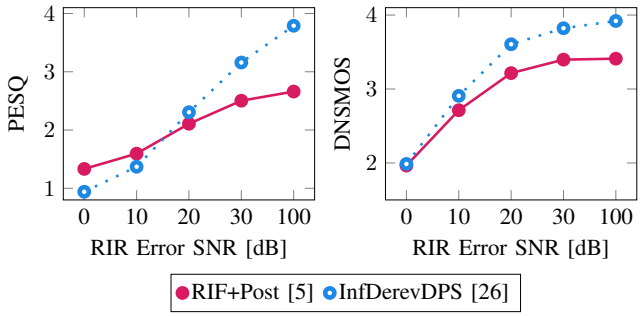


Fig. 1: *Robustness of informed dereverberation approachess with respect to normally distributed errors in the RIR.*

TABLE I: *Dereverberation results on matched reverberant VCTK dataset. We indicate for each method in the table if it operates in a blind scenario.*

Method	Blind	DNS-MOS	PESQ
Reverberant	-	3.14 ± 0.52	1.61 ± 0.37
RIF+Post [5]	✗	3.41 ± 0.47	2.66 ± 0.40
InfDerevDPS [26]	✗	3.91 ± 0.33	3.95 ± 0.42
FiNS/RIF+Post [5], [41]	✓	2.18 ± 0.38	1.33 ± 0.19
FiNS/InfDerevDPS [26], [41]	✓	2.19 ± 0.43	1.32 ± 0.18
BUDDy (ours) [40]	✓	**3.76 ± 0.41**	**2.30 ± 0.53**

in the provided RIR. This is problematic since perfect sample-wise estimation of RIRs is an arduous problem, given the statistical nature of RIRs [11].

We now shift to a more practical scenario where the RIR is estimated from the reverberant speech only by a DNN-based estimator. In particular, we employ FiNS [41], a state-of-the-art supervised RIR estimator which obtains RIR estimates based on the reverberant utterance (see Section V-C1 for details). We compare in Table I results where the RIR is perfectly known (i.e. informed scenario) versus when it is estimated by FiNS (i.e. partially blind). The acoustic conditions in the considered evaluation set match those of the training set. Therefore, since FiNS was trained in a supervised fashion using paired reverberant/RIR data, it is expected to perform well on such conditions. Indeed, through informal listening, we notice that FiNS produces perceptually reasonable RIR estimates. Yet, the dereverberation performance of both InfDerevDPS and RIF+Post is poor when the RIR is estimated with FiNS [41], as opposed to when the RIR is perfectly known. This shows the limited robustness of these informed methods, and suggests that in such blind case the RIR should be jointly estimated with the anechoic speech. This is the paradigm followed by our method BUDDy, which we will introduce in the next sections.

III. Informed Diffusion-Based Dereverberation

This section introduces diffusion models, a class of generative models that form the foundation of the proposed method. It also explores their application in solving inverse problems, specifically highlighting their use in informed dereverberation.

A. Diffusion-Based Generative Models

Diffusion models [22], [42] have achieved remarkable success across various domains, including speech [43]. They break down the problem of generating high-dimensional complex data into a series of easier denoising tasks. Training a diffusion model first requires defining a *forward process*, which gradually adds noise to data points, turning the target data distribution into a tractable Gaussian distribution. Conversely, data generation is accomplished by reversing the corruption process. First, an initial sample is drawn from a Gaussian distribution, and then the model iteratively removes noise until a clean sample from the target distribution emerges.

The *reverse process*, which defines a transport between a Gaussian prior distribution and a target data distribution p_{data}, can be characterized by the *probability flow* ordinary differential equation (ODE):

$$\mathrm{d}\mathbf{x}_\tau = [\mathbf{f}(\mathbf{x}_\tau, \tau) - \frac{1}{2} g^2(\tau) \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)]\mathrm{d}\tau\,, \tag{1}$$

where diffusion time τ flows in reverse from $\tau = T$ to $\tau = T_{\min} \ll T$. The diffusion state $\mathbf{x}_\tau \in \mathbb{R}^L$ starts from the initial condition $\mathbf{x}_T \in \mathbb{R}^L$ and ends at $\mathbf{x}_0 \in \mathbb{R}^L \sim p_{\text{data}}$, where L is the length of the time-domain speech utterance. We adopt the parameterization proposed by Karras et al. [44], which defines the *drift* and *diffusion* parameters as $f(\mathbf{x}_\tau, \tau) = 0$ and $g(\tau) = \sqrt{2\tau}$, respectively. Similarly, we adopt $\sigma(\tau) = \tau$ as the noise schedule which defines the so-called *transition kernel* i.e. the marginal density of the forward process:

$$q_\tau(\mathbf{x}_\tau | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0, \sigma^2(\tau)\mathbf{I})\,, \tag{2}$$

where $\mathbf{I} \in \mathbb{R}^{L \times L}$ is the identity matrix. The *score function* $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)$ indicates the direction of maximum data likelihood. In practice, it is intractable and we need to estimate it with a *score model* $\mathbf{s}_\theta(\mathbf{x}_\tau, \tau)$ parameterized with a deep neural network (DNN). Vincent et al. have shown that the score model $\mathbf{s}_\theta(\mathbf{x}_\tau, \tau)$ can be optimized using denoising score matching, i.e. matching the score of the Gaussian transition kernel $q_\tau(\mathbf{x}_\tau | \mathbf{x}_0)$ instead of the score of the unknown probability $p(\mathbf{x}_\tau)$ [45]. The score of the transition kernel $q_\tau(\mathbf{x}_\tau | \mathbf{x}_0)$ can be obtained from (2) as:

$$\nabla_{\mathbf{x}_\tau} q_\tau(\mathbf{x}_\tau | \mathbf{x}_0) = \frac{\mathbf{x}_\tau - \boldsymbol{\mu}(\mathbf{x}_0, \tau)}{\sigma(\tau)^2}\,. \tag{3}$$

The score model $\mathbf{s}_\theta$ is therefore trained using the denoising score-matching objective: [45]

$$\mathbb{E}_{\substack{\tau \sim \mathcal{U}(T_{\min}, T_{\max}) \\ \mathbf{x}_0 \sim p_{\text{data}} \\ \mathbf{x}_\tau \sim q_\tau(\mathbf{x}_\tau | \mathbf{x}_0)}} \left[\lambda(\tau) \left\| \mathbf{s}_\theta(\mathbf{x}_\tau, \tau) - \frac{\mathbf{x}_\tau - \boldsymbol{\mu}(\mathbf{x}_0, \tau)}{\sigma^2(\tau)} \right\|_2^2 \right], \tag{4}$$

where first a diffusion index τ is randomly sampled between extremal times $T_{\min}$ and $T_{\max} > T$, a data point $\mathbf{x}_0$ is sampled in the training set, and the corresponding diffusion state $\mathbf{x}_\tau$ is obtained from the transition kernel in (2). In practice, we use the same pre-conditioning for $\mathbf{s}_\theta(\mathbf{x}_\tau, \tau)$ and same loss weighting $\lambda(\cdot)$ as in Karras et al. (see [44] for details).

Algorithm 1 Reverberation Operator $\mathcal{A}_\psi(\cdot)$

Require: parameters ψ, speech estimate $\hat{\mathbf{x}}_0$
$\{\mathbf{\Phi}, (w_b, \alpha_b)_{b=1,\dots,B}\} \leftarrow \psi$ ▷ Parameter set
$\mathbf{A}'_{n,b} \leftarrow w_b \cdot e^{-\alpha_b n}$ ▷ Exponential decay model
$\mathbf{A} \leftarrow \exp\{\text{lerp}(\log \mathbf{A}')\}$ ▷ Frequency interpolation
$\mathbf{H} = \mathbf{A} e^{j\mathbf{\Phi}}$
$\overline{\mathbf{H}} = \text{STFT}(\delta_\text{d} \oplus \mathcal{P}_\text{min}(\text{iSTFT}(\mathbf{H})))$ ▷ Projection step
$\hat{\mathbf{X}} = \text{STFT}(\hat{\mathbf{x}}_0)$
$\hat{\mathbf{Y}}_{m,k} \leftarrow \sum_{n=0}^{N_h} \overline{\mathbf{H}}_{n,k} \hat{\mathbf{X}}_{m-n,k}$ ▷ Subband convolution
return $\text{iSTFT}(\hat{\mathbf{Y}})$

B. Diffusion Posterior Sampling for Dereverberation

We discuss in this section how diffusion priors can be adapted in order to solve inverse problems. While some traditional methods derive maximum a posteriori estimators for blind dereverberation [9]–[11], we exploit the generative nature of diffusion models to solve this inverse problem using posterior sampling. Assuming for now that the RIR $\mathbf{h}$ is known, we attempt to sample from the posterior distribution of the anechoic speech given the measurement and the RIR $p(\mathbf{x}_0|\mathbf{y}, \mathbf{h})$. This is achieved by solving the probability flow ODE (1), replacing the score function $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)$ by the *posterior score* $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau|\mathbf{y}, \mathbf{h})$ [23]. The posterior score is obtained through Bayes' rule as:

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau|\mathbf{y}, \mathbf{h}) = \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau) + \nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau, \mathbf{h}) \,. \quad (5)$$

The first term, or *prior score*, is directly obtained via the score model $\mathbf{s}_\theta(\mathbf{x}_\tau, \tau)$. The likelihood score $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau, \mathbf{h})$ is in general intractable for $\tau > 0$. As in [27], we employ an estimate of $\mathbf{x}_0$, denoted as $\hat{\mathbf{x}}_0$, and we assume that this estimate is a sufficient statistic for $\mathbf{x}_\tau$. This results in a first assumption $p(\mathbf{y}|\mathbf{x}_\tau, \mathbf{h}) \approx p(\mathbf{y}|\hat{\mathbf{x}}_0, \mathbf{h})$. The estimate $\hat{\mathbf{x}}_0(\mathbf{x}_\tau)$ is obtained as the posterior mean of $\mathbf{x}_0$ knowing $\mathbf{x}_\tau$ and is derived using Tweedie's formula, i.e. one-step denoising of $\mathbf{x}_\tau$:

$$\hat{\mathbf{x}}_0(\mathbf{x}_\tau) \triangleq \mathbb{E}[\mathbf{x}_0|\mathbf{x}_\tau] \approx \mathbf{x}_\tau - \sigma^2(\tau)\mathbf{s}_\theta(\mathbf{x}_\tau, \tau). \quad (6)$$

In order to approximate $p(\mathbf{y}|\hat{\mathbf{x}}_0, \mathbf{h})$, previous work [26] models the error between $\mathbf{y}$ and $\mathbf{h} * \hat{\mathbf{x}}_0$ to follow a zero-mean Gaussian distribution in the time domain. The corresponding expression for the likelihood score $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau, \mathbf{h})$ is then a simple weighted L^2-distance between $\mathbf{y}$ and $\mathbf{h} * \hat{\mathbf{x}}_0$. However we observed that far better dereverberation performance and speech quality can be achieved by substituting the obtained distance with a L^2-distance between compressed spectrograms instead. This is analogous to modelling the likelihood score as:

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau, \mathbf{h}) \approx -\zeta(\tau) \nabla_{\mathbf{x}_\tau} \mathcal{C}(\mathbf{y}, \mathbf{h} * \hat{\mathbf{x}}_0(\mathbf{x}_\tau)) \,, \quad (7)$$

where $\zeta(\tau)$ is a diffusion-time-dependent scaling parameter that controls the influence of the likelihood score term in the sampling trajectory, and $\mathcal{C}(\cdot, \cdot)$ is the following cost function:

$$\mathcal{C}(\mathbf{u}, \mathbf{v}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \| S_\text{comp}(\mathbf{u})_{m,k} - S_\text{comp}(\mathbf{v})_{m,k} \|_2^2. \quad (8)$$

There, $S_\text{comp}(\mathbf{u}) = |\text{STFT}(\mathbf{u})|^{2/3} \exp\{j \angle \text{STFT}(\mathbf{u})\}$ is the magnitude-compressed spectrogram of $\mathbf{u}$. This cost function implies that we model the reconstruction error in the compressed STFT domain as a Gaussian with unknown variance $\frac{1}{2\zeta(\tau)}$. We apply this compression to boost low-energy components as typically observed in high frequencies of speech signals or in late reverberation tails, and account for the heavy-tailedness of speech distributions [4]. Such a strategy is also employed in [46] for data representation.

The parameter $\zeta(\tau)$ balances a trade-off between adherence to the prior data distribution and fidelity to the observed data. We empirically resort to the same parameterization of $\zeta(\tau)$ as in [32], [47]:

$$\zeta(\tau) = \frac{\sqrt{L}\,\tilde{\zeta}}{\| \nabla_{\mathbf{x}_\tau} \mathcal{C}\left(\mathbf{y}, \mathbf{h} * \hat{\mathbf{x}}_0(\mathbf{x}_\tau)\right) \|_2} \,, \quad (9)$$

where $\tilde{\zeta}$ is a fixed coefficient.

Algorithm 2 Inference algorithm

Require: Reverberant speech $\mathbf{y}$
$\mathbf{x}_\text{init} \leftarrow \text{WPE}(\mathbf{y})$
Sample $\mathbf{x}_N \sim \mathcal{N}(\mathbf{x}_\text{init}, \sigma_N^2 \mathbf{I})$ ▷ Warm initialization
Initialize ψ_N ▷ Initialize the RIR parameters
for $n \leftarrow N, \dots, 1$ **do** ▷ Discrete step backwards
$\mathbf{s}_n \leftarrow s_\theta(\mathbf{x}_n, \tau_n)$ ▷ Evaluate score model
$\hat{\mathbf{x}}_0 \leftarrow \mathbf{x}_n - \sigma_n^2 \mathbf{s}_n$ ▷ Get one-step denoising estimate
$\hat{\mathbf{x}}_0 \leftarrow \text{Rescale}(\hat{\mathbf{x}}_0)$ ▷ Constraint RMS power
$\psi_{n-1}^0 \leftarrow \psi_n$ ▷ Use RIR parameters from last step
for $j \leftarrow 0, \dots, N_\text{its.}$ **do** ▷ RIR optimization
$\mathcal{J}_\text{RIR}(\psi_{n-1}^j) \leftarrow \mathcal{C}(\mathbf{y}, \mathcal{A}_{\psi_{n-1}^j}(\hat{\mathbf{x}}_0)) + \mathcal{R}(\psi_{n-1}^j)$
$\psi_{n-1}^{j+1} \leftarrow \psi_{n-1}^j - \text{Adam}(\mathcal{J}_\text{RIR}(\psi_{n-1}^j))$ ▷ Opti. step
$\psi_{n-1}^{j+1} \leftarrow \text{clamp}(\psi_{n-1}^{j+1})$ ▷ Constrain Parameters
$\psi_{n-1} \leftarrow \psi_{n-1}^M$
$\mathbf{g}_n \leftarrow -\zeta(\tau_n) \nabla_{\mathbf{x}_n} \mathcal{C}(\mathbf{y}, \mathcal{A}_{\psi_{n-1}}(\hat{\mathbf{x}}_0))$ ▷ LH score approx.
$\mathbf{x}_{n-1} \leftarrow \mathbf{x}_n - \sigma_n(\sigma_{n-1} - \sigma_n)(\mathbf{s}_n + \mathbf{g}_n)$ ▷ Update step
return $\mathbf{x}_0$ ▷ Reconstructed audio signal

IV. Blind Diffusion-Based Dereverberation

This section elaborates on the proposed method for blind dereverberation, where the impulse response $\mathbf{h}$ is unknown. In Section IV-A, we define a reverberation operator $\mathcal{A}_\psi(\cdot)$, which comprises a structured parametric model of the RIR, with parameters ψ. This operator is summarized in Algorithm 1. The proposed inference method BUDDy, detailed in IV-B, performs joint speech dereverberation and RIR estimation by combining the conditional sampling from a diffusion model with an optimization of the RIR parameters. The complete inference procedure is summarized in Algorithm 2, and the processing pipeline is visualized in Fig.2.

A. Reverberation Operator

1) Subband Filtering: In contrast to Section III, here we model reverberation using a subband filtering approximation in the short-time Fourier transform (STFT) domain [48], [49]. Let $\mathbf{H} := \text{STFT}(\mathbf{h}) \in \mathbb{C}^{N_\mathbf{h} \times K}$ represent the STFT of a RIR $\mathbf{h}$ with $N_\mathbf{h}$ time frames and K frequency bins. Similarly,

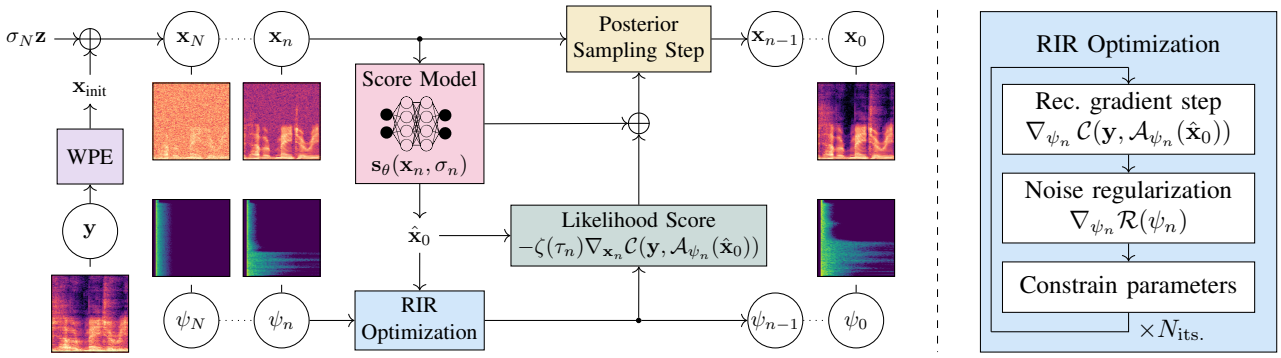


Fig. 2: *BUDDy: joint optimization alternating between RIR estimation and posterior sampling for speech reconstruction [40].*

let $\mathbf{X} \in \mathbb{C}^{M \times K}$, and $\mathbf{Y}$, denote the STFTs of anechoic $\mathbf{x}_0$ and reverberant $\mathbf{y}$ speech signals, repectively. The subband convolution operation applies independent convolutions along the time dimension of each frequency band:

$$\mathbf{Y}_{m,k} = \sum_{n=0}^{N_h} \mathbf{H}_{n,k} \mathbf{X}_{m-n,k} \,. \tag{10}$$

The resulting reverberant signal $\mathbf{Y} \in \mathbb{C}^{(M+N_\mathbf{h}-1)\times K}$ can be transformed to time domain by applying the inverse STFT. The subband filtering model only approximates the time-domain convolution, as it does not account for the spectral leakage between frequency bands. However, it is empirically found to be a valid assumption in many scenarios involving reverberation [7], [49], [50]. In our case, we noticed that adding 50% zero-padding to the end of the frames before computing the STFT was important to avoid cyclic convolution artifacts when retrieving the resulting signal to the time domain.

2) Room Impulse Response Prior: In the blind scenario, we need to estimate $\mathbf{H}$, which is an ill-posed problem task when not knowing the anechoic speech. Therefore, we need to constrain the space of possible solutions by imposing a prior on $\mathbf{H}$. We propose a structured, differentiable prior on $\mathbf{H}$, whose parameters ψ can be estimated through gradient descent. We denote the complete forward reverberation operator, including forward and inverse STFT operations, as $\mathcal{A}_\psi(\cdot) : \mathbb{R}^L \to \mathbb{R}^{L+L_\mathbf{h}-1}$. The whole processing pipeline is summarized in Algorithm 1 with each component detailed below.

We denote as $\mathbf{A} \in \mathbb{R}^{N_\mathbf{h} \times K}$ and $\mathbf{\Phi} \in \mathbb{R}^{N_\mathbf{h} \times K}$ the RIR magnitudes and phases, respectively. Following [11], we adopt an exponential decay model with learnable parameters controlling the decay time. Since room materials exhibit frequency-dependent absorption behavior, we parameterize the magnitude matrix $\mathbf{A}$ as a multi-band exponential decay model defined in $B < K$ frequency bands. Let $\mathbf{A}' \in \mathbb{R}^{N_\mathbf{h} \times B}$ be the subsampled version of $\mathbf{A}$ in the B selected frequency bands. Each frequency band b is characterized by its weight w_b and exponential decay rate α_b, such that the corresponding subband magnitude filter is derived as

$$\mathbf{A}'_{n,b} = w_b \cdot e^{-\alpha_b n} \,. \tag{11}$$

Note that our parameterization can be extended to model coupled spaces by employing several decay parameters per band and summing their respective contributions [51]. We found it beneficial to constrain w_b and α_b within a limited range to stabilize the optimization, specially at early stages. This is achieved by clamping the parameters to predefined minimum and maximum values after every optimization iteration, as specified in Appendix C1c. Once the parameters are estimated, we reconstruct the K-bands magnitudes $\mathbf{A}$ by interpolating the subsampled $\mathbf{A}'$ as $\mathbf{A} = \exp(\text{lerp}(\log(\mathbf{A}')))$, where lerp represents linear interpolation on the frequency scale. For this purpose, we employ the `torchcde` library, which facilitates efficient and differentiable interpolation [52]. After interpolation of the magnitude matrix, we then obtain the time-frequency RIR $\mathbf{H}$ by multiplying the magnitude matrix $\mathbf{A}$ with the complex phase exponentials:

$$\mathbf{H} = \mathbf{A} \odot e^{j\mathbf{\Phi}}, \tag{12}$$

where j is the imaginary number and $\odot$ represents element-wise multiplication. Given the general lack of phase structure, we optimize each phase factor in $\mathbf{\Phi}$ independently. The RIR model $\psi = \{\mathbf{\Phi}, (w_b, \alpha_b)_{b=1,\dots,B}\}$ ultimately contains $2 \times B + N_\mathbf{h} \times K$ optimizable parameters.

3) Projections: We extend our forward reverberation operator with a series of projections to increase the likelihood of generating plausible RIRs. Thus, the time-frequency RIR $\mathbf{H}$ is further processed as:

$$\overline{\mathbf{H}} = \text{STFT}\left(\delta \oplus \mathcal{P}_{\text{min}}(\text{iSTFT}(\mathbf{H}))\right) \,. \tag{13}$$

This primarily ensures STFT consistency of $\overline{\mathbf{H}}$, exploiting the redundancy of the STFT representation and imposing inter-frame correlations between the RIR phases Φ. We then enforce that the time-domain RIR estimate $\mathbf{h}$ has minimum-phase lag, using the Hilbert transform-based method in [53]. This is indicated by the operator $\mathcal{P}_{\text{min}}$ and guarantees stability of the inverse RIR filter [2]. We refer the reader to Appendix A for further details. Finally, the operation $\delta \oplus (\cdot)$ replaces the first sample of the time-domain RIR with a unit impulse. This has the effect of injecting knowledge of the direct path in $\overline{\mathbf{H}}$, and further requires us to correct the magnitude matrix $\mathbf{A}$ to account for this operation. It is important to note that these steps are integral to the reverberation operator $\mathcal{A}_\psi(\cdot)$, which maps the parameters ψ to the convolved signal $\mathcal{A}_\psi(\hat{\mathbf{x}}_0)$, as outlined in Algorithm 1. Since all operations are differentiable,

we compute gradients with respect to ψ by backpropagating through all operations. We propose a detailed ablation study of these projection and correction steps in Section V-A6.

B. Blind Dereverberation Inference

We aim to solve the following joint dereverberation and RIR parameter optimization problem:

$$\hat{\mathbf{x}}_0, \hat{\psi} = \underset{\mathbf{x}_0, \psi}{\arg\min}\, \mathcal{C}(\mathbf{y}, \mathcal{A}_\psi(\mathbf{x}_0)) + \mathcal{R}(\psi) \;\text{ s.t. }\; \mathbf{x}_0 \sim p_{\text{data}} \quad (14)$$

where $\mathcal{C}(\mathbf{y}, \mathcal{A}_\psi(\mathbf{x}_0))$ is the reconstruction error with $\mathcal{C}$ the cost function introduced in (8), and $\mathcal{R}(\psi)$ is a RIR prior regularization term. This objective seeks to find the optimal speech $\hat{\mathbf{x}}_0$ and RIR parameters $\hat{\psi}$ that minimize both losses while imposing the soft constraint that the estimated signal $\hat{\mathbf{x}}_0$ should adhere to the anechoic speech distribution p_{data}. We leverage the pre-trained score model $\mathbf{s}_\theta(\mathbf{x}_\tau, \tau)$ trained on anechoic speech to enforce this constraint.

We optimize (14) by solving the following ODE, obtained from the classical probability-flow ODE (1) where we injected the specified diffusion parameters of Karras et al. [44] and the likelihood score approximation derived in (7):

$$\mathrm{d}\mathbf{x}_\tau = -\tau \left[\mathbf{s}_\theta(\mathbf{x}_\tau, \tau) - \zeta(\tau) \nabla_{\mathbf{x}_\tau} \mathcal{C}(\mathbf{y}, \mathcal{A}_{\psi_\tau}(\hat{\mathbf{x}}_0))\right] \mathrm{d}\tau\,. \quad (15)$$

The regularization term $\mathcal{R}(\psi)$ introduced in (14) is:

$$\mathcal{R}(\psi) = \frac{1}{N_{\mathbf{h}}} \sum_{m=1}^{N_{\mathbf{h}}} \sum_{k=1}^{K} \| S_{\text{comp}}(\hat{\mathbf{h}}_\psi)_{m,k} - S_{\text{comp}}(\hat{\mathbf{h}}_{\psi'} + \sigma' \mathbf{v})_{m,k} \|_2^2, \quad (16)$$

where $\hat{\mathbf{h}}_\psi = \mathcal{A}_\psi(\delta)$ is the current time-domain RIR estimate and $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a vector of white Gaussian noise. The noise level schedule $\sigma'(\tau)$ is similar to the score model schedule $\sigma(\tau)$ but its values are limited to avoid having too much noise at early steps (see Appendix C1a) The term $\hat{\mathbf{h}}_{\psi'}$ represents a copy of $\hat{\mathbf{h}}_\psi$, such that the $\arg\min$ in (14) does not apply to it.. In other terms, we detach the gradients of $\hat{\mathbf{h}}_\psi$ from the optimization graph to obtain $\hat{\mathbf{h}}_{\psi'}$, similar to what is done in e.g. [54] for guiding data reconstruction. We provide in Appendix B a short analysis of the regularization objective $\mathcal{R}(\psi)$. We show that this term injects multiplicative noise with standard deviation $\sigma'(\tau)$ in the RIR parameter gradients, which arguably smoothes the RIR parameter search.

During optimization, we further rescale the denoised speech estimate $\hat{\mathbf{x}}_0$ so that its root-mean-square power (RMS) matches the average RMS power of clean speech computed on the training set. Using this additional constraint helps lifting the indeterminacy when jointly optimizing the speech $\mathbf{x}_0$ and RIR parameters ψ. This stesp is included in our ablation study in Section V-A6.

In order to guide and accelerate reverse diffusion, it is often beneficial to use warm initialization, i.e. to let the reverse diffusion process start from a speech sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_{\text{init}}, \sigma^2(T)\mathbf{I})$, where $\mathbf{x}_{\text{init}}$ has some interesting cues about the clean signal we wish to estimate. Similar to [34], we obtain the initial mean signal $\mathbf{x}_{\text{init}}$ through WPE [7], a blind dereverberation algorithm based on variance-normalized delayed linear prediction. WPE performs mild dereverberation, which allows us to get closer to the clean speech, while not introducing too many distortions to the signal. As WPE is blind and unsupervised, our method remains fully blind and unsupervised as well.

V. Experiments and Results

In this section, we provide a comprehensive evaluation of BUDDy across various datasets and experimental setups. We detail the methodologies and baselines employed and present the results of our experiments.

A. Speech Dereverberation

We present dereverberation results on 16kHz speech data, building upon the experiments conducted in prior work [40].

1) Data: We use VCTK [57] as clean speech, selecting 103 speakers for training, 2 for validation and 2 for testing. The total dataset represents 44h of audio, which we down-sample to 16kHz for our experiments. We curate RIRs from various public datasets [58]–[66]. In total we approximately obtain 10k RIRs, and split them between training, validation and testing using ratios 0.9/0.05/0.05. We also generate another RIR dataset for testing methods in a mismatched setting, using RIRs simulated with `pyroomacoustics` [67]. For ease of comparison, we choose simulation parameters such that the distributions of reverberation times and direct-to-reverberation ratios of the simulated mismatched dataset approximately match those of the real-recorded matched dataset.

2) Baselines: We compare our method BUDDy to several blind supervised baselines such as the predictive approach in [55], which will denote as PSE in the following (for *predictive speech enhancement*), and diffusion-based SGMSE+ [24] and StoRM [25]. The STFT-based diffusion model in SGMSE+ and StoRM uses supervision in both the network conditioning and the diffusion trajectory parameterization; PSE uses a classical L^2-distance between the clean target and its estimate and has virtually the same architecture as SGMSE+. These methods require coupled reverberant/anechoic speech, which we generate using our curated RIR and anechoic speech datasets. The reverberant speech is obtained by first aligning the direct path of the RIR to its first sample, then convolving the anechoic speech from VCTK with the resulting RIR, and finally normalizing it to reach the same loudness [68] as the anechoic speech.

We also include blind unsupervised approaches leveraging traditional methods such as WPE [7] and Yohena and Yatabe [8], as well as generative models Saito et al. [34], GibbsDDRM [35] and RVAE-EM [38]. Please see Appendix C1b for more details on baselines.

3) Hyperparameters: As in [26], [40], we implement the unconditional score model architecture with NCSN++M [25], [55], which is a convolution-based neural network operating in the complex STFT domain. NCSN++M is also used as the base architecture for PSE, SGMSE+ and StoRM. Details on the architecture, training configuration, reverberation operator and diffusion hyperparameters can be found in appendices C1a., C1c and C1d, respectively.

TABLE II: *Speech dereverberation results on reverberant VCTK datasets. We indicate for each method in the table whether it is supervised or not. Boldface numbers indicate best performance for supervised and unsupervised methods separately.*

		Matched			Mismatched		
Method	Unsup.	DNS-MOS	PESQ	ESTOI	DNS-MOS	PESQ	ESTOI
Reverberant	-	3.14 ± 0.52	1.61 ± 0.37	0.50 ± 0.14	3.05 ± 0.47	1.57 ± 0.29	0.47 ± 0.11
PSE	✗	3.75 ± 0.38	2.85 ± 0.55	0.80 ± 0.10	3.61 ± 0.39	2.08 ± 0.47	0.64 ± 0.09
SGMSE+M [24], [55]	✗	3.88 ± 0.32	2.99 ± 0.48	0.78 ± 0.09	3.74 ± 0.34	2.48 ± 0.47	**0.69 ± 0.09**
StoRM [25]	✗	**3.90 ± 0.33**	**3.33 ± 0.48**	**0.82 ± 0.10**	**3.83 ± 0.32**	**2.51 ± 0.53**	0.67 ± 0.09
Yohena and Yatabe [8]	✓	2.99 ± 0.56	1.80 ± 0.33	0.55 ± 0.12	2.94 ± 0.44	1.71 ± 0.29	0.51 ± 0.10
WPE [56]	✓	3.24 ± 0.54	1.81 ± 0.42	0.57 ± 0.14	3.10 ± 0.48	1.74 ± 0.37	0.54 ± 0.12
Saito et al. [34]	✓	3.22 ± 0.56	1.68 ± 0.40	0.51 ± 0.13	3.12 ± 0.52	1.70 ± 0.33	0.52 ± 0.10
GibbsDDRM [35]	✓	3.33 ± 0.53	1.70 ± 0.37	0.51 ± 0.13	3.30 ± 0.52	1.75 ± 0.36	0.52 ± 0.11
RVAE-EM [38]	✓	3.05 ± 0.53	1.83 ± 0.32	0.54 ± 0.11	3.00 ± 0.45	1.76 ± 0.30	0.52 ± 0.10
BUDDy (ours)	✓	**3.76 ± 0.41**	**2.30 ± 0.53**	**0.66 ± 0.12**	**3.74 ± 0.38**	**2.24 ± 0.54**	**0.65 ± 0.12**

4) Instrumental metrics: For instrumental evaluation of the speech dereverberation performance, we use the intrusive Perceptual Evaluation of Speech Quality (PESQ) [69] and extended short-term objective intelligibility (ESTOI) [70] for assessment of speech quality and intelligibility respectively. We also use the non-intrusive DNS-MOS [71], a DNN-based mean opinion score (MOS) approximation following the ITU-T P.835 recommendation [72].

5) Instrumental evaluation results: We display in Table II the dereverberation results for all blind methods, both supervised and unsupervised. Blind supervised approaches-PSE, SGMSE+ and StoRM generally perform better than unsupervised methods as they benefit from supervision at training time. However, we can observe the limited generalization ability of supervised approaches in the setting where acoustic conditions are mismatched, i.e. when using simulated RIRs. Our method BUDDy, however, seamlessly adapts to changing acoustics since it was trained without supervision. This enables BUDDy to keep the same performance on both *matched* and *mismatched* datasets, where supervised methods like PSE lose as much as 0.77 PESQ points in mismatched conditions. Furthermore, BUDDy performs far better than all other blind unsupervised baselines. For instance, BUDDy outperforms WPE by as much as 0.50 PESQ and 0.10 ESTOI points. Indeed, traditional unsupervised methods [7], [8] only draw limited benefits from their uninformed Gaussian prior on anechoic speech, while diffusion-based Saito et al. [34] and GibbsDDRM [35] seem to only marginally deviate from their WPE initialization. RVAE-EM [38] also obtains low instrumental scores, but informal listening suggested that its dereverberation abilities were superior to those of WPE.

6) Ablation study: We conduct an ablation study to evaluate the impact of the projection step (13) introduced in the operator optimization (see Section IV-A). We present the results in Table III and observe that, although the minimum-phase consistency projection has a theoretical justification as a mean to enhance the stability of the inverse RIR during optimization, its practical effect appears negligible. However, we observe that the other operations in the projection step, i.e. STFT consistency, enforcement of the direct path, and speech magnitude constraint, are all instrumental in guiding BUDDy toward a solution with higher fidelity to clean speech, as measured by PESQ. We show DNS-MOS figures out of completeness. However, DNS-MOS variations are small across ablations and not indicative of fidelity to reference speech as DNS-MOS is not an intrusive metric.

Additionally, we examine the effect of parameterizing the likelihood model with a L^2-distance on compressed spectrograms rather than on waveforms as in previous work [26]. To do so, we replace the cost function $C(\cdot, \cdot)$ from (8), which is based on compressed spectrogams, with a simpler waveform-domain L^2-distance. The results clearly show the superiority of using a cost function on compressed spectrograms.

7) Listening experiment: Instrumental metrics offer only limited insights into the performance of dereverberation algorithms [73]. We therefore conduct a listening experiment based on the MUSHRA recommendation [74] to assess the performance of BUDDy as perceived by human listeners. The test comprised 12 pages, featuring 6 reverberant speech utterances from each of the *matched* and *mismatched* datasets. Participants were asked to rate the different stimuli with a single number representing overall quality, taking into account factors such as voice distortion, residual reverberation, and potential artifacts [73]. The test stimuli include our proposed method BUDDy, the unsupervised WPE [7] and RVAE-EM [38], as well as the supervised baselines PSE and SGMSE [24], Further details on the organization of the listening experiment are reported in Appendix C1e.

The results of the experiment are presented in Fig. 3. It can be observed that the unsupervised baselines WPE and RVAE-EM received low scores. Yet, RVAE-EM performs consistently better than WPE in this listening experiment, as opposed to what is suggested by instrumental metrics in Table II. In the matched test set (Fig. 3a), BUDDy obtained significantly lower scores than PSE and SGMSE+ ($p < 0.001$ in a paired Welch test). However, in the mismatched set, PSE and SGMSE+ suffered a decrease in performance, losing up to 20 points (out of 100), while BUDDy mantained similar scores. In that case, there is no significant difference in performance between

Fig. 3: *Listening test results on reverberant VCTK datasets. The boxplot shows first quartile, median and third quartile.*

the three approaches ($p > 0.1$), which closes the gap between BUDDy and the top-performing supervised baselines in this mismatched setting, highlighting the advantage provided by unsupervised learning.

B. Singing Voice Dereverberation

We extend our evaluation benchmark to include the related task of singing voice dereverberation.

1) Data: We collect several publicly available singing voice datasets [75]–[80]. These datasets feature over 94 h of studio-quality solo singing from a diverse array of singers and singing styles, spanning various languages. The majority of the recordings are in Chinese, followed by English, Japanese, and Korean. All datasets are down-sampled to 44.1 kHz. For testing, similar to [35], we use the sung part of NHSS [81], [82]. The NHSS dataset contains 100 English-language pop songs, 10 for each of the five male and five female singers recruited. We select a subset (90%) of the RIRs curated for the VCTK-based experiments, such that we only retain the RIRs whose original sample rate is at least 44.1kHz.

2) Baselines: We evaluate the performance of BUDDy against two unsupervised baselines: WPE [7] and the unsupervised method from Saito et al. [34] which was originally designed for singing voice dereverberation. Due to the lack of established supervised baselines for this specific task, we created our own by adapting the diffusion model from BUDDy to a supervised setting. In this adaptation, we trained a diffusion model on dry singing voice using the same architecture and hyperparameters as the unconditional model in BUDDy, but added a paired sample of reverberant voice as a condition. The conditioning sample is incorporated into the architecture by stacking it in the channel dimension. This setup resembles the "variance exploding" approach used by Gonzalez et al. for speech enhancement [83]. We train all methods on our singing voice dataset detailed in the previous section.

3) Hyperparameters: As in [32] for music restoration, we use the UNet architecture proposed in [84] without self-attention blocks, and wrap the computations within an invertible Constant-Q Transform (CQT) [85]. The CQT produces a time-frequency representation where pitch transpositions are equivalent to frequency-wise translations, highlighting its interest for processing signals with harmonic components such as singing voice and music. More details with regard to the architecture and specific training configuration and inference hyperparameters are reported in Appendix C2

TABLE III: *Ablation study on reverberant VCTK dataset.*

Method	PESQ	DNS-MOS
Reverberant	1.61 ± 0.37	3.14 ± 0.52
BUDDy	**2.30 ± 0.53**	3.76 ± 0.41
- Minimum-phase Consistency	**2.30 ± 0.57**	3.81 ± 0.40
- RMS Power Constraint	2.22 ± 0.50	3.64 ± 0.50
- Fixed Direct Path	2.10 ± 0.46	3.78 ± 0.44
- STFT Consistency	1.96 ± 0.41	**3.84 ± 0.39**
L^2-Distance for $\mathcal{C}(\cdot,\cdot)$	1.86 ± 0.47	3.36 ± 0.56

TABLE IV: *Singing voice dereverberation results on reverberant NHSS dataset. We indicate each method whether it is unsupervised or not.*

Method	Unsup.	ℓ^1 STFT	FAD (VGGish)
Reverberant	-	1.98 ± 0.66	6.41
Conditional Diffusion	✗	**1.56 ± 0.48**	1.71
WPE [7]	✓	2.02 ± 0.65	4.53
Saito et al. [34]	✓	1.95 ± 0.65	5.41
BUDDy (ours)	✓	1.92 ± 0.60	**1.32**

4) Evaluation metrics: Objective metrics for evaluating singing voice restoration tasks are currently more limited compared to those available for speech processing. Following [34], we use the ℓ^1-distance in the magnitude STFT domain and a Fréchet Audio Distance (FAD) using a VGGish embedding [86]. However, these are only limited in interpretability and hardly relate to listening impression [87], [88]. Therefore, we complete this evaluation benchmark with a listening test with 10 participants, using a similar setup as reported in Section V-A7. In this case, the test included 10 reverberant singing voice examples from the NHSS dataset, and the instructions were identical to those reported in Appendix C1e.

5) Results: The results from the instrumental evaluation are reported in Table IV and those from the listening test in Fig. 4. The results show that BUDDy largely outperforms the unsupervised baselines and is on-par with the conditional diffusion model that benefited from supervision at training time. We found no statistically significant difference between the listening test scores of BUDDy and the conditional diffusion model (p-value of 0.55 in a paired Welch test).

C. Room Impulse Response Estimation

BUDDy is not only designed as a dereverberation algorithm but also functions as a blind unsupervised RIR estimator. We evaluate its performance for RIR estimation using the same speech model and data we employed for speech dereverberation in Section V-A.

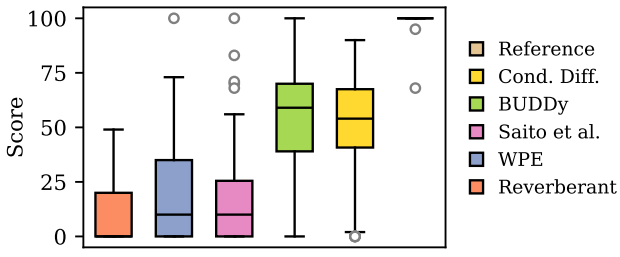


Fig. 4: *Listening test results on singing voice dataset. The boxplot shows first quartile, median and third quartile.*

1) Baseline: We benchmark BUDDy against FiNS [41], a deep neural network (DNN)-based approach trained to estimate time-domain RIRs directly from reverberant speech. FiNS comprises a 1D-convolutional encoder and a two-component decoder. The first decoder component estimates the late tail of the RIR by passing noise signals through a trainable filterbank containing several FIR filters. The second decoder component directly estimates the direct path and early reflections in the time-domain. In contrast to BUDDy, FiNS relies on supervised learning, thus requiring a paired dataset of reverberant speech and RIRs. We use an unofficial re-implementation[2] and train the model on our VCTK-based reverberant speech dataset.

2) Evaluation metrics: Due to the highly ill-posed nature of the blind RIR estimation problem and the statistical nature of late reflections [11], we refrain from using element-wise distances, such as error-to-signal ratios, to evaluate the performance of RIR estimators. Instead, it is arguably more important to preserve the acoustic and perceptual properties of the reference RIR [89]. On the other hand, single metrics such as the T_{60} reverberation time or clarity index C_{50} do not account for frequency-specific estimation errors. We therefore employ a subband reverberation time T_{60} and clarity index C_{50}, with subbands spanning octaves. This enables to keep a high-level representation of the acoustical properties while allowing enough granularity on the spectral attributes of the RIR. The reverberation time T_{60} is defined for a diffuse sound field as the time it takes for its energy decay curve (EDC) to decay by 60dB [1]. To compute T_{60} from a RIR while avoiding the effects of the noise floor, we first determine T_{30} and extrapolate it to T_{60} by multiplying by a factor of 2. We calculate T_{30} as the time required for the EDC to decrease from -5 dB to -35 dB relative to the initial level to eliminate the influence of the direct path. This measure is computed in each octave band separately. The octave clarity index C_{50} is the ratio (in dB) between the energy in the 50 first milliseconds and the energy in the remaining of the RIR, calculated in the corresponding octave band [1]. Consequently, we compute the absolute error between the T_{60} and C_{50} values calculated for each octave from the estimated RIR and those from the ground truth RIR.

[2]https://github.com/kyungyunlee/fins

3) Results: The results for both matched and mismatched test sets are plotted in Fig. 5. In the matched condition, FiNS and BUDDy achieve similar T_{60} error rates at low- and mid-range frequency bands, while BUDDy's performance decreases at high frequencies (Fig. 5a). Our intuition is that the lower RIR estimation abilities of BUDDy at high frequencies can be linked to the tendency of diffusion models to generate high-frequency components at the later stages of the reverse diffusion process [90]. Consequently, there is less information available for optimizing the RIR parameters in this range, negatively affecting parameter convergence. A similar trend is observed for the C_{50} error (Fig. 5c). Furthermore, BUDDy generally achieves lower C_{50} error than FiNS in the mid-frequency range, where most of the speech content lies.

In the mismatched setting, FiNS struggles to generalize because of its supervised training setup. As a result, BUDDy outperforms FiNS in both T_{60} and C_{50} error at low and mid-frequency bands (Figs. 5b) and 5d). At higher frequencies, BUDDy's T_{60} estimation performance still remains slightly inferior to FiNS, though the gap is noticeably smaller than in the matched setting. Regarding C_{50}, BUDDy outperforms FiNS on all frequency domains but the higher, 4-kHz-centered band. This increased relative performance of BUDDy compared to FiNS highlights the benefits of leveraging unsupervised training for RIR estimation in variable acoustic conditions.

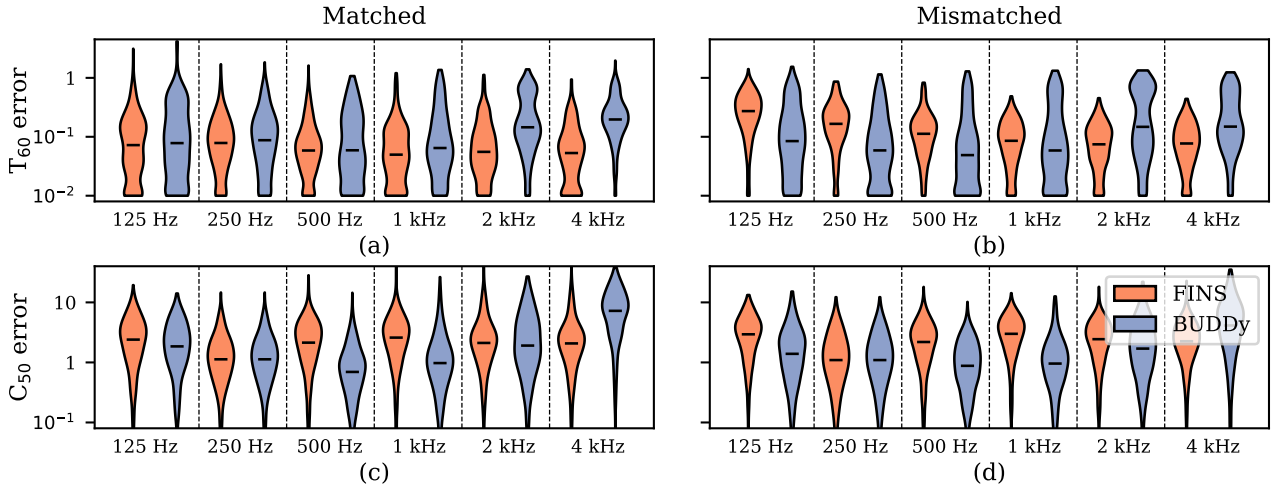


Fig. 5: *RIR estimation metrics calculated for each octave on the reverberant VCTK dataset. The violin plots show the results distribution, with the median highlighted. Lower values indicate better performance. FiNS [41] is trained in a supervision fashion whereas BUDDy is unsupervised.*

VI. Conclusion

In this paper, we presented an unsupervised method that simultaneously performs blind dereverberation and RIR estimation using diffusion models. Our results highlight the importance of joint speech and RIR estimation in contrast to plugging estimated RIRs into informed dereverberation methods. The proposed method, BUDDy, yields state-of-the-art performance among unsupervised approaches for blind speech and singing voice dereverberation, outperforming both traditional and diffusion-based methods. Unlike blind supervised methods, which often struggle with generalization to unseen acoustic conditions, our unsupervised approach naturally overcomes this limitation due to its ability to adapt the reverberation operator to a broad range of room impulse responses. This holds as well for RIR estimation, as we show that the RIR estimation performance of BUDDy surpasses that of a state-of-the-art supervised DNN-based technique in mismatched acoustic conditions while being on par in a matched setting.

Acknowledgements

We would like to thank Koichi Saito, Fumikuri Yohena and Kohei Yatabe for providing us with code and guidance through their methods.

References

[1] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, vol. 59, Springer, 2011.

[2] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE TASLP*, vol. 36, no. 2, pp. 145–152, 1988.

[3] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. John Wiley & Sons, 2018.

[4] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proc. IWAENC*, 2010.

[5] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2014.

[6] J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 1982.

[7] T. Nakatani, B. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model," *IEEE TASLP*, vol. 16, no. 8, pp. 1512–1527, 2008.

[8] F. Yohena and K. Yatabe, "Single-channel blind dereverberation based on rank-1 matrix lifting in time-frequency domain," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2024.

[9] D. Schmid, S. Malik, and G. Enzner, "A maximum a posteriori approach to multichannel speech dereverberation and denoising," in *Proc. IWAENC*, 2012.

[10] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Speech dereverberation with convolutive transfer function approximation using map and variational deconvolution approaches," in *Proc. IWAENC*, 2014.

[11] E. A. P. Habets, *Speech Dereverberation Using Statistical Reverberation Models*, pp. 57–93, Springer, London, 2010.

[12] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, no. 1, pp. 165–169, 1979.

[13] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP J. Adv. Signal Process.*, 2007.

[14] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE TASLP*, vol. 26, no. 10, pp. 1702–1726, 2018.

[15] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM TASLP*, vol. 25, no. 7, pp. 1492–1501, 2017.

[16] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM TASLP*, vol. 23, no. 6, pp. 982–992, 2015.

[17] X. Liu, S.-J. Chen, and J. H. Hansen, "Dual-path minimum-phase and all-pass decomposition network for single channel speech dereverberation," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2024.

[18] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *Proc. EUSIPCO*, 2019.

[19] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM TASLP*, vol. 28, pp. 1598–1607, 2020.

[20] J.-M. Lemercier, J. Richter, S. Welker, E. Moliner, V. Välimäki, and T. Gerkmann, "Diffusion models for audio restoration," *arXiv*, 2024.

[21] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. ICML*, 2015.

[22] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, 2020.

[23] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. ICLR*, 2021.

[24] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM TASLP*, vol. 31, pp. 2351–2364, 2023.

[25] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE TASLP*, vol. 31, pp. 2724–2737, 2023.

[26] J.-M. Lemercier, S. Welker, and T. Gerkmann, "Diffusion posterior sampling for informed single-channel dereverberation," in *Proc. WASPAA*, 2023.

[27] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," in *Proc. ICLR*, 2023.

[28] H. Chung, J. Kim, S. Kim, and J. C. Ye, "Parallel diffusion models of operator and image for blind inverse problems," in *Proc. CVPR*, 2023.

[29] C. Laroche, A. Almansa, and E. Coupeté, "Fast diffusion em: a diffusion model for blind inverse problems with application to deconvolution," *IEEE/CVF WACV*, 2024.

[30] Y. Sanghvi, Y. Chi, and S. H. Chan, "Kernel diffusion: An alternate approach to blind deconvolution," *arXiv*, 2023.

[31] B. Nortier, M. Sadeghi, and R. Serizel, "Unsupervised speech enhancement with diffusion-based generative models," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2024.

[32] E. Moliner, F. Elvander, and V. Välimäki, "Blind audio bandwidth extension: A diffusion-based zero-shot approach," *arXiv*, 2024.

[33] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Proc. NeurIPS*, 2022.

[34] K. Saito, N. Murata, T. Uesaka, C.-H. Lai, Y. Takida, T. Fukui, and Y. Mitsufuji, "Unsupervised vocal dereverberation with diffusion-based generative models," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2023.

[35] N. Murata, K. Saito, C.-H. Lai, Y. Takida, T. Uesaka, Y. Mitsufuji, and S. Ermon, "GibbsDDRM: A partially collapsed Gibbs sampler for solving blind inverse problems with denoising diffusion restoration," in *Proc. ICML*, 2023.

[36] H. Attias, J. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Proc. NeurIPS*, 2000.

[37] D. Baby and H. Bourlard, "Speech dereverberation using variational autoencoders," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2021.

[38] P. Wang and X. Li, "RVAE-EM: Generative speech dereverberation based on recurrent variational auto-encoder and convolutive transfer function," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2024.

[39] Z.-Q. Wang, "USDNet: Unsupervised speech dereverberation via neural forward filtering," 2024.

[40] E. Moliner, J.-M. Lemercier, S. Welker, T. Gerkmann, and V. Välimäki, "BUDDy: Single-channel blind unsupervised dereverberation with diffusion models," in *Proc. IWAENC*, 2024.

[41] C. J. Steinmetz, V. K. Ithapu, and P. Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," in *Proc. WASPAA*, 2021.

[42] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. NeurIPS*, 2019.

[43] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," *Proc. ICLR*, 2021.

[44] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Proc. NeurIPS*, 2022.

[45] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[46] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Proc. Interspeech*, 2022.

[47] E. Moliner, J. Lehtinen, and V. Välimäki, "Solving audio inverse problems with a diffusion model," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2023.

[48] M. M. Goodwin, "Realization of arbitrary filters in the stft domain," in *Proc. WASPAA*, 2009.

[49] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE TASLP*, vol. 15, no. 4, pp. 1305–1319, 2007.

[50] J.-M. Lemercier, J. Tobergte, and T. Gerkmann, "Extending DNN-based multiplicative masking to deep subband filtering for improved dereverberation," in *Proc. Interspeech*, 2023.

[51] N. Xiang and P. M. Goggans, "Evaluation of decay times in coupled spaces: Bayesian decay model selection," *J. Acoust. Soc. Am.*, vol. 113, no. 5, pp. 2685–2697, 2003.

[52] P. Kidger, J. Morrill, J. Foster, and T. Lyons, "Neural Controlled Differential Equations for Irregular Time Series," *Proc. NeurIPS*, 2020.

[53] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice—Hall, 1975.

[54] M. Mardani, J. Song, J. Kautz, and A. Vahdat, "A variational perspective on solving inverse problems with diffusion models," in *Proc. ICLR*, 2024.

[55] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing discriminative versus diffusion generative models for speech restoration tasks," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2023.

[56] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2008.

[57] C. Valentini-Botinhao et al., "Reverberant speech database for training speech dereverberation algorithms and TTS models," *University of Edinburgh*, 2016.

[58] K. Prawda, S. J. Schlecht, and V. Välimäki, "Calibrating the Sabine and Eyring formulas," *J. Acoust. Soc. Am.*, vol. 152, no. 2, pp. 1158–1169, 2022.

[59] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE challenge—Corpus description and performance evaluation," in *Proc. WASPAA*, 2015.

[60] M. Jeub, M. Schaefer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. Int. Conf. Dig. Signal Process.*, 2009.

[61] D. Fejgin, W. Middelberg, and S. Doclo, "Brudex database: Binaural room impulse responses with uniformly distributed external microphones," in *Proc. ITG Conf. Speech Communication*, 2023.

[62] U. of Kent, "Palimpsest impulse responses," https://research.kent.ac.uk/sonic-palimpsest/impulse-responses.

[63] G. Kearney et al., "Measuring the acoustical properties of the BBC Maida Vale recording studios for virtual reality," *Acoustics*, vol. 4, no. 3, pp. 783–799, 2022.

[64] B. U. of Technology, "BUT speech@FIT reverb database," https://speech.fit.vutbr.cz/software/but-speech-fit-reverb-database.

[65] T. Dietzen, R. A. Ali, M. Taseska, and T. van Waterschoot, "MYRiAD: A multi-array room acoustic database," *EURASIP J. Audio Speech and Music Process.*, , no. 17, pp. 1–14, 2023.

[66] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *National Academy of Sciences*, vol. 113, no. 48, pp. 7856–7865, 2016.

[67] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2018.

[68] G. International Telecommunication Union, "Algorithms to measure audio programme loudness and true-peak audio level," Rec. BS.1770-4, Geneva, Switzerland, Oct. 20023.

[69] 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 *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2001.

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

[71] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *arXiv*, 2021.

[72] G. International Telecommunication Union, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," Rec. P.835, Geneva, Switzerland, Oct. 2003.

[73] S. Goetze, A. Warzybok, I. Kodrasi, J. O. Jungmann, B. Cauchi, J. Rennies, E. A. P. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, "A study on speech quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms," in *Proc. IWAENC*, 2014.

[74] G. International Telecommunication Union, "Method for the subjective assessment of intermediate quality level of audio systems," Rec. BS.1534-3, Geneva, Switzerland, Oct. 2015.

[75] R. Huang et al., "Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus," in *ACM International Conference on Multimedia*, 2021.

[76] Y. Wang et al., "Opencpop: A high-quality open source Chinese popular song corpus for singing voice synthesis," *arXiv*, 2022.

[77] L. Zhang et al., "M4singer: A multi-style, multi-singer and musical score provided Mandarin singing corpus," in *Proc. NeurIPS*, 2022.

[78] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *Proc. APSIPA ASC*, 2013.

[79] S. Choi, W. Kim, S. Park, S. Yong, and J. Nam, "Children's song dataset for singing voice research," in *Proc. ISMIR*, 2020.

[80] J. Koguchi, S. Takamichi, and M. Morise, "PJS: Phoneme-balanced Japanese singing-voice corpus," in *Proc. APSIPA ASC*, 2020.

[81] B. Sharma, X. Gao, K. Vijayan, X. Tian, and H. Li, "NHSS: A speech and singing parallel database," *arXiv*, 2020.

[82] B. Sharma and H. Li, "A combination of model-based and feature-based strategy for speech-to-singing alignment," in *Proc. Interspeech*, 2019.

[83] P. Gonzalez, Z.-H. Tan, J. Østergaard, J. Jensen, T. S. Alstrøm, and T. May, "Investigating the design space of diffusion models for speech enhancement," *arXiv*, 2023.

[84] E. Moliner and V. Välimäki, "Diffusion-based audio inpainting," *J. Audio Eng. Soc.*, vol. 72, pp. 100–113, Mar. 2024.

[85] G. A. Velasco, N. Holighaus, M. Dörfler, and T. Grill, "Constructing an invertible constant-Q transform with non-stationary Gabor frames," in *Proc. DAFX*, 2011.

[86] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," *arXiv preprint arXiv:1812.08466*, 2018.

[87] N. Kandpal, O. Niteo, and Z. Jin, "Music enhancement via image translation and vocoding," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2022.

[88] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, "Adapting Frechet audio distance for generative music evaluation," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2024, pp. 1331–1335.

[89] G. D. Santo, K. Prawda, S. J. Schlecht, and V. Välimäki, "Similarity metrics for late reverberation," in *Proc. Asilomar Conf. Signal Sys. Comp.*, 2024.

[90] X. Yang, D. Zhou, J. Feng, and X. Wang, "Diffusion probabilistic model made slim," in *Proc. CVPR*, 2023.

[91] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. ICLR*, 2015.

APPENDIX

A. Minimum phase constraint

The minimum-phase constraint in Section IV-A takes the time-domain RIR $\mathbf{h}$ and computes the minimum-delay phase Θ as:

$$\Theta = -\operatorname{Im}\left[\mathcal{H}\left(\log\left|\mathcal{F}(\mathbf{h})\right|\right)\right], \tag{17}$$

where $\mathcal{F}$ is the Fourier transform and $\mathcal{H}$ the Hilbert transform:

$$\mathcal{H}(\mathbf{x}) \triangleq \mathcal{F}^{-1}(-j \cdot \operatorname{sign}(\omega)\mathcal{F}(\mathbf{x})) \tag{18}$$

The minimum-delay corrected time-domain RIR is then obtained by replacing the original phase with the obtained minimum-delay phase:

$$\mathbf{h}_{\text{min}} = \mathcal{F}^{-1}(|\mathcal{F}(\mathbf{h})|e^{j\Theta}) \tag{19}$$

It is worth noting that all the operations involved in this method are differentiable, which allows backpropagation throughout the process.

B. Noise regularization

We introduce in Section IV-B a noise regularization term, which we can write in a simplified fashion, ignoring the sums and indexations, as:

$$\mathcal{R}(\psi) = \|S_{\text{comp}}(\hat{\mathbf{h}}_\psi) - S_{\text{comp}}(\hat{\mathbf{h}}_{\psi'} + \sigma'\mathbf{v})\|_2^2, \tag{20}$$

The gradient computed during optimization is obtained as:

$$\begin{aligned}\frac{\partial \mathcal{R}(\psi)}{\partial \psi} &= 2\left(S_{\text{comp}}(\hat{\mathbf{h}}_\psi) - S_{\text{comp}}(\hat{\mathbf{h}}_{\psi'} + \sigma'\mathbf{v})\right) \\ &\quad \times \frac{\partial S_{\text{comp}}}{\partial \hat{\mathbf{h}}_\psi} \times \left(\frac{\partial \hat{\mathbf{h}}_\psi}{\partial \psi} - \underbrace{\frac{\partial \hat{\mathbf{h}}_{\psi'}}{\partial \psi}}_{0}\right) \\ &\approx -2\sigma'\mathbf{v}\left[\frac{\partial S_{\text{comp}}}{\partial \hat{\mathbf{h}}_\psi}\right]^2 \frac{\partial \hat{\mathbf{h}}_\psi}{\partial \psi}\end{aligned}$$

where we have ignored second- and higher-order Taylor expansion terms of S_{comp} for simplicity. We observe that the resulting gradient for $\mathcal{R}(\psi)$ is proportional to the noise vector $\mathbf{v}$ and to the gradient of the estimated RIR $\hat{\mathbf{h}}(\psi)$ with respect to the parameters ψ. Therefore, adding $\mathcal{R}(\psi)$ in the optimization has the result of adding multiplicative noise to the operator gradients (with respect to ψ) which emerge from the optimization of the reconstruction loss $\mathcal{C}(\mathbf{y}, \mathcal{A}_\psi(\mathbf{x}_0))$.

Empirically, this has the effect of smoothing out the optimization of the RIR operator parameters ψ and avoiding degenerate solutions, provided that the dedicated noise schedule σ' is reasonably chosen.

C. Experimental details

1) Speech Dereverberation:

a) Architecture and training hyperparameters: We train the unconditional score model $\mathbf{s}_\theta$ for our method BUDDy with anechoic data only, using segments of 4 seconds randomly extracted from the utterances in VCTK. Same as in [26], [40], we implement the unconditional score network architecture with NCSN++M [25], [55], a lighter variant of the NCSN++ [23] with 27.8M parameters. Similar to [47], we wrap up the network with a time-frequency transform, in this case the STFT, such that the NCSN++M forward pass is effectively performed in the complex STFT domain using a real and imaginary parts representation. For all methods, STFTs are computed using a Hann window of 32 ms and a hop size of 8 ms. The complex prediction at every state can be converted to time-domain by inverting the STFT. We adopt Adam [91] as the optimizer to train the unconditional score model, with a learning rate of 10^{-4} and an effective batch size of 16 for 200k iterations. We track an exponential moving average of the DNN weights with a decay of 0.999 to be used for sampling as in [24].

b) Baselines: For WPE [7], we take 5 iterations, a filter length of 50 STFT frames (400ms) and a delay of 2 STFT frames (16ms). We set the hyperparameters of the method by Yohena and Yatabe [8] to $M = 50$ and $\rho = 400$ after conducting a parameter search. Using code gently provided by the authors, we retrain Saito et al. [34] and GibbsDDRM [35] using the same data as for BUDDy, i.e. the anechoic VCTK dataset. We use the same inference parameters which can be found in [34], [35] although we tried to improve the results by doing a hyperparameter search as suggested by the authors. We re-train RVAE-EM in unsupervised mode on our anechoic VCTK dataset using publicly available code and use the original inference parameters reported by the authors [38].

c) Reverberation operator: The STFT parameters are the same as those used in the unconditional score model, i.e. we use a Hann window of 32 ms and a hop size of 8 ms. For subband filtering we further employ 50% zero-padding to avoid frequency aliasing artifacts. Given our sampling rate of $f_s = 16$ kHz, this results in $K = 513$ unique frequency bins. We set the number of STFT frames of our operator to $N_{\mathbf{h}} = 100$ (800ms). We subsample the frequency scale in $B = 26$ bands, with a 125 Hz spacing between 0 and 1kHz, a 250Hz spacing between 1 and 2kHz, and a 500Hz spacing between 3 and 8kHz.

We optimize the RIR parameters ψ using Adam, with a learning rate of 0.1, and the momentum parameters are set to $\beta_1 = 0.9$, and $\beta_2 = 0.99$. We employ $N_{\text{its.}} = 10$ optimization iterations per diffusion step. We further constrain the weights w_b between 0 and 40 dB, and the decays α_b between 0.5 and 28. This avoids the optimization from approaching degenerate solutions, especially at the early stages of sampling.

d) Forward and reverse diffusion: As mentioned in Section IV-B we obtain our initial estimate $\mathbf{x}_{\text{init}}$ through WPE dereverberation. Consequently, we choose $T = 0.5$ such that the initial noise in $\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_{\text{init}}, \sigma^2(T)\mathbf{I})$ effectively masks potential artifacts stemming from WPE, while still retaining the general structure in $\mathbf{x}_{\text{init}}$ that may guide the process. We set the minimal diffusion time to $T_{\text{min}} = 10^{-4}$ and adopt the same reverse discretization scheme as Karras et al. [44]:

$$\forall i < N,\ \tau_i = \sigma_i = \left(T^{1/\rho} + \tfrac{i}{N-1}(T_{\text{min}}^{1/\rho} - T^{1/\rho})\right)^\rho, \tag{21}$$

with warping $\rho = 10$ and $N = 200$ steps. We use the second-order Euler-Heun stochastic sampler in [44] with $S_{\text{churn}} = 50$. In the noise regularization term depicted in (16), the annealing schedule σ' follows the same discretization as σ, but we restrict its values between $\sigma'_{\text{min}} = 5 \times 10^{-4}$ and $\sigma'_{\text{max}} = 10^{-2}$. The scaling factor used for the variance estimate $\eta(\tau)$ in (9) is fixed to $\tilde{\eta} = 0.6$.

e) Listening experiment: We conduct a listening experiment based on the MUSHRA recommendation [74] using

the webMUSHRA[3] interface. The test comprised 12 pages, featuring 6 reverberant speech utterances from each of the *matched* and *mismatched* datasets. The test was conducted in isolated conditions within listening booths at the Aalto Acoustics Lab. In total, 10 volunteers participated in the experiment. All utterances were loudness-normalized to -23dB LUFS. The participants were allowed to modify the volume of headphones during the training stage (first page, not included in the results). The ground-truth anechoic speech served as the reference, which was also hidden among the other conditions (WPE, RVAE-EM, PSE, SGMSE, BUDDy), while the original reverberant speech signal was used as the low anchor, expected to receive a score of 0. Participants were advised to focus particularly on dereverberation performance and to use the full rating scale, i.e. rate reference as 100 and reverberant anchor as 0. We obtained consent directly from the participants through a written form. As the study did not present any risk for the subjects, no review board was required for the approval of this experiment.

2) Singing Voice Dereverberation:

We use the UNet architecture proposed in [84] without self-attention blocks, and wrap the computations within an invertible Constant-Q Transform (CQT) [85]. The resulting architecture consists of 45M parameters. We employ a 1534-point window and hop size 384 for the CQT. The unconditional score model is optimized using Adam with same parameters as for the VCTK dataset, but we reduce the batch size to 4 and use 6-second anechoic audio segments. For this experiment, we use $B = 39$ bands for the subband decomposition in the reverberation operator for BUDDy, extending the bands used in Appendix C1c above 8kHz with a 1kHz spacing.

[3]https://github.com/audiolabs/webMUSHRA

A.6 HRTF Estimation using a Score-based Prior [P13]

Abstract

We present a head-related transfer function (HRTF) estimation method which relies on a data-driven prior given by a score-based diffusion model. The HRTF is estimated in reverberant environments using natural excitation signals, e.g. human speech. The impulse response of the room is estimated along with the HRTF by optimizing a parametric model of reverberation based on the statistical behaviour of room acoustics. The posterior distribution of HRTF given the reverberant measurement and excitation signal is modelled using the score-based HRTF prior and a log-likelihood approximation. We show that the resulting method outperforms several baselines, including an oracle recommender system that assigns the optimal HRTF in our training set based on the smallest distance to the true HRTF at the given direction of arrival. In particular, we show that the diffusion prior can account for the large variability of high-frequency content in HRTFs.

Reference

Étienne Thuillier, Jean-Marie Lemercier, Eloi Moliner, Timo Gerkmann and Vesa Välimäki, "HRTF Estimation using a Score-based Prior", *Preprint. Submitted to: IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP).*, 2024, DOI: 10.48550/arXiv.2410.01562

Copyright Notice

Authors' Contributions

Étienne Thuillier originated the idea of transposing the unsupervised BUDDy model for HRTF estimation. He implemented the recommender baselines, trained the unconditional HRTF prior, prepared the HRTF and binaural data, conducted preliminary experiments that contributed to the development of the model, wrote Sections I, V and co-authored Section III. Jean-Marie Lemercier is the second author of this publication. He ran several ablations helping develop the model, and ran the final evaluations. He co-authored Section III. He wrote Sections II, IV and VI. He wrote the abstract. Eloi Moliner implemented most parts of the algorithm, and ran a significant part of the experiments for developing the model. The first three authors revised together the manuscript. Timo Gerkmann and Vesa Välimäki brought insights on the methods and mathematical derivations.

HRTF Estimation using a Score-based Prior

Etienne Thuillier
Acoustics Lab
Aalto University
Espoo, Finland

Jean-Marie Lemercier
Signal Processing Group
University of Hamburg
Hamburg, Germany

Eloi Moliner
Acoustics Lab
Aalto University
Espoo, Finland

Timo Gerkmann
Signal Processing Group
University of Hamburg
Hamburg, Germany

Vesa Välimäki
Acoustics Lab
Aalto University
Espoo, Finland

***Abstract*—We present a head-related transfer function (HRTF) estimation method which relies on a data-driven prior given by a score-based diffusion model. The HRTF is estimated in reverberant environments using natural excitation signals, e.g. human speech. The impulse response of the room is estimated along with the HRTF by optimizing a parametric model of reverberation based on the statistical behaviour of room acoustics. The posterior distribution of HRTF given the reverberant measurement and excitation signal is modelled using the score-based HRTF prior and a log-likelihood approximation. We show that the resulting method outperforms several baselines, including an oracle recommender system that assigns the optimal HRTF in our training set based on the smallest distance to the true HRTF at the given direction of arrival. In particular, we show that the diffusion prior can account for the large variability of high-frequency content in HRTFs.**

***Index Terms*—3D audio, diffusion models, head-related transfer function, spatial audio.**

I. Introduction

Standard head-related transfer function (HRTF) measurements follow a system identification approach whereby a synthetic probe signal is rendered through a loudspeaker, picked-up by microphones located at the entrances of the subject's ear canals, and used to deconvolve the resulting binaural recording [1]. Repeating the measurement to capture a full HRTF is laborious and time-consuming. The procedure also requires a dedicated anechoic chamber and specialized, calibrated, audio equipment. Methods using non-specialized equipment in echoic environments have been proposed to democratize access to individualized HRTFs, for example, using a living room's loudspeaker emitting short bursts of exponential sine sweeps [2] or a hand-held smartphone emitting synthetic probe signals in near field [3]. While scalable to the mass-market and cost-effective, these approaches require the subjects to actively undergo a procedure during which they are subjected to unpleasant-sounding synthetic signals.

This issue could be addressed by relying on isolated sound sources occurring in the subject's surrounding environment. Recently, Jayaram et al. [4] trained a neural network in a supervised fashion to predict the HRTF's magnitude spectrum using recordings captured from consumer-grade binaural microphones. This method relies on detection of the sound source's location and composes a full HRTF by aggregating estimates from various direction of arrivals (DoAs). Such an approach could prove particularly advantageous if shown adaptable to modern earbud headphones, in which microphones are typically offset from the entrances of the ear canals.

In this work, we also propose to leverage binaural recordings of a source in the subject's everyday environment. However, we suggest using sounds played back by the user over a paired device with known directivity patterns, for example podcast content over a smart speaker, such that the source is known a-priori along with its DoA. This setup leads us to formulating the task of HRTF estimation from a blind inverse problem perspective, i.e. sampling a valid HRTF that is consistent with the observed binaural reverberant measurement, while also estimating the reverberation in the room. The HRTF sampling procedure uses a data-driven prior provided by a diffusion model trained on binaural time-aligned HRTF filter data. We then fit the room acoustics to a parametric model adapted from [5], [6], which we jointly optimize during the HRTF estimation. This follows a recent line of work that applies diffusion models as priors to solve inverse problems in image and audio domains [6]–[8].

Unlike [4], the proposed approach recovers both magnitude and phase estimates. In contrast to a previous signal processing method under a similar setup [9], the prior ensures consistent scaling between measurements. Furthermore, we demonstrate that our approach outperforms a nearest-neighbour oracle baseline which returns the HRTF from the training set that is closest to the true HRTF at the the detected DoA. In particular, the diffusion prior shows a good expressivity in the higher-frequency regions where the HRTF presents large variations across subjects. The size of the diffusion model is modest, potentially allowing for on-device processing.

II. Score-based Diffusion Model of HRTFs

We present here our data-driven prior for time-aligned HRTF features based on continuous-time diffusion-based generative models, also known as score-based models. Score-based models [10], [11] encompass a class of generative models particularly successful at learning complex data distributions such as e.g. natural images or human speech. Here, we employ score-based models to approximate $p(\mathbf{a}|\gamma)$, that is the distribution of time-aligned HRTFs in the frequency domain, denoted as $\mathbf{a} \in \mathbb{C}^{2\times F}$, conditioned on the DoA γ.

Score-based models operate as iterative Gaussian denoisers: during training, the target data distribution is transformed into a standard Gaussian distribution following a *forward*

diffusion process, incrementally adding noise. Once training is achieved, new data belonging to the data distribution can be generated through the *reverse diffusion process*, which iteratively removes noise from an initial Gaussian sample until a data sample emerges. In continuous-time score-based models [11], this reverse process can be characterized by the following *probability flow* ordinary differential equation (ODE) [12], adopting the parameterization by Karras et al. [13]:

$$\mathrm{d}\mathbf{a}_\tau = -\tau \nabla_{\mathbf{a}_\tau} \log p(\mathbf{a}_\tau|\gamma)\mathrm{d}\tau, \tag{1}$$

where τ indexes the reverse process flowing from $T_{\max}$ to 0. The diffusion state $\mathbf{a}_\tau$ starts from the initial condition $\mathbf{a}_{T_{\max}} \sim \mathcal{N}(0, \sigma(T_{\max})^2\mathbf{I})$ and terminates at $\mathbf{a}_0 \sim p_{\text{data}}$. We choose a linear noise variance schedule $\sigma(\tau) = \tau$, which defines the Gaussian marginal densities $p_\tau(\mathbf{a}_\tau|\mathbf{a}_0) = \mathcal{N}(\mathbf{a}_\tau; \mathbf{a}_0, \sigma(\tau)^2\mathbf{I})$. The *score* $\nabla_{\mathbf{a}_\tau} \log p(\mathbf{a}_\tau|\gamma)$ is intractable at inference for complex distributions. Therefore, a *score model* parameterized with a deep neural network (DNN) $\mathbf{s}_\theta(\mathbf{a}_\tau, \tau, \gamma)$ is trained to estimate the score using *denoising score matching* [14].

III. Binaural Room Impulse Response Parameterization

In a static setup where source and receiver locations are fixed, binaural reverberation can be modelled by convolving an anechoic source $\mathbf{s}$ with the impulse response of the system, i.e. the binaural room impulse response (BRIR) $\mathbf{h}$. The BRIR is composed from the contributions of wavefronts traveling from the source to the ears of the subject following direct and indirect propagation paths. In this work, we model the BRIR as the sum of an anechoic and a reverberant component:

$$\mathbf{h}_{\mathbf{a}}^{(\psi)} := \begin{bmatrix} \boldsymbol{\delta}_{t_{\text{left}}} \\ \boldsymbol{\delta}_{t_{\text{left}}} \end{bmatrix} * \left(g \begin{bmatrix} \boldsymbol{\delta}_0 \\ \boldsymbol{\delta}_{t_{\text{itd}}} \end{bmatrix} * \mathcal{F}^{-1}(\mathbf{a}) + \mathbf{r}^{(\chi)} \right), \tag{2}$$

where $\psi = \{t_{\text{left}}, g, t_{\text{itd}}\} \cup \chi$ denotes the optimizable parameters of the model, and $\mathbf{a}$ is the binaural time-aligned HRTF [15] at the given DoA. The HRTF is defined in the frequency domain, hence its time-domain equivalent called head-related impulse response is obtained via inverse Fourier transform $\mathcal{F}^{-1}$. The symbol $\boldsymbol{\delta}_t$ denotes a Kronecker delta delayed by t seconds, $t_{\text{left}} \in [0, \infty)$ models time of flight to the subject's left ear, $g \in (0, 1]$ represents a real-valued attenuation constant, $t_{\text{itd}} \in \mathbb{R}$ denotes the inter-aural time difference (ITD) taking the left ear as the reference (non-delayed) channel, and $\mathbf{r} \in \mathbb{R}^{2\times N_r}$ denotes the reverberant component of the model.

The reverberant component $\mathbf{r}$ is implemented as a binaural variant of a previously proposed parametric model fitting the reverberant characteristics of the room [5], [6]:

$$\mathbf{r}^{(\chi)} := \mathcal{F}_{ST}^{-1} \left(\text{interp.} \left(\left[w_b \, \mathrm{e}^{-m\alpha_b} \right]_{(c,m,b)} \right) \odot \left[\mathrm{e}^{j\phi_{c,m,f}} \right]_{(c,m,f)} \right), \tag{3}$$

where $\chi = \left\{ [\alpha_b]_b, [w_b]_b, [\phi_{c,m,f}]_{c,m,f} \right\}$ denotes the learnable parameters of $\mathbf{r}$, $\mathcal{F}_{ST}^{-1}$ denotes the inverse short-time Fourier transform (STFT) , $\odot$ denotes the Hadamard element-wise product, $m \in \{1, \ldots, M_{\mathbf{r}}\}$ indexes over STFT frames, $f \in \{1, \ldots, F\}$ over Fourier bins, $b \in \{1, \ldots, B\}$ over frequency subbands (with $B < F$), and $c \in \{\text{left}, \text{right}\}$.

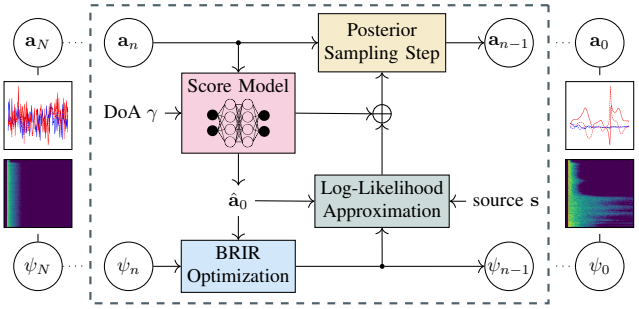


Fig. 1: Diagram of the inference algorithm.

The weight $w_b \in [0, \infty)$ and decay rate $\alpha_b \in [0, \infty)$ define a subband exponential decaying magnitude model, which exploits the observed statistical nature of reverberation tails [16]. Crucially, this magnitude envelope is identical for the left and right channels. The interpolation operator interp. upsamples the B subbands to the F frequency bins of the STFT by employing $\exp(\text{lerp}(\log(\cdot)))$, where lerp denotes linear interpolation. In contrast to the magnitude envelope term, the phase is determined using a specific coefficient $\phi_{c,m,f} \in [\pi, \pi)$ for each channel-frame-bin triplet. This allows for fitting decorrelated left and right channel realizations of the diffuse part, as typically observed in BRIRs above 1 kHz [17], [18]. Additionally, while the early reflections of the BRIR are not explicitly incorporated into the model, the unconstrained phases allow the model to fit these reflections to some extent.

IV. Inference Algorithm

The inference process solves the following objective:

$$\hat{\mathbf{a}}, \hat{\psi} = \arg\min_{\mathbf{a}, \psi} \mathcal{C}(\mathbf{y}, \mathbf{h}_{\mathbf{a}}^{(\psi)} * \mathbf{s}) \quad \text{s.t.} \quad \mathbf{a} \sim p_{\text{data}} \tag{4}$$

This means that we wish to retrieve the optimal time-aligned HRTF $\hat{\mathbf{a}}$ and BRIR parameters $\hat{\psi}$ in order to minimize a reconstruction error $\mathcal{C}(\mathbf{y}, \mathbf{h}_{\mathbf{a}}^{(\psi)} * \mathbf{s})$ given and the binaural reverberant measurement $\mathbf{y}$ and the broadband excitation signal $\mathbf{s}$, which is in our case (but not restricted to) human speech. Another soft constraint is that the estimated time-aligned HRTF $\hat{\mathbf{a}}$ should belong to the target HRTF distribution p_{data}.

We solve (4) using an alternating optimization procedure, visualized in Fig. 1. The acoustic parameters $\hat{\psi}$ in our BRIR model (3) are updated with a classical gradient-based optimizer (e.g. Adam) minimizing the reconstruction loss. The HRTF estimate $\hat{\mathbf{a}}$, however, is refined using a posterior sampling technique leveraging a score model $\mathbf{s}_\theta(\mathbf{a}_\tau, \tau, \gamma)$ trained on time-aligned HRTF data and conditioned on the DoA γ. Precisely, this procedure aims at sampling from the posterior distribution $p(\mathbf{a}|\mathbf{y}, \mathbf{s}, \gamma)$. We leverage our HRTF score-based prior for posterior sampling by solving the ODE (1), where the score function $\nabla_{\mathbf{a}_\tau} \log p(\mathbf{a}_\tau|\gamma)$ is replaced by the so-called *posterior score* obtained via Bayes' formula:

$$\nabla_{\mathbf{a}_\tau} \log p(\mathbf{a}_\tau|\mathbf{y}, \mathbf{s}, \gamma) = \nabla_{\mathbf{a}_\tau} \log p(\mathbf{a}_\tau|\gamma) + \nabla_{\mathbf{a}_\tau} \log p(\mathbf{y}|\mathbf{a}_\tau, \mathbf{s}) \,. \tag{5}$$

The *prior score* $\nabla_{\mathbf{a}_\tau} \log p(\mathbf{a}_\tau|\gamma)$ is obtained via our score model $\mathbf{s}_\theta(\mathbf{a}_\tau, \tau, \gamma)$. Since we generally do not have a model for $\mathbf{y}$ given the diffusion state $\mathbf{a}_\tau$, the likelihood $p(\mathbf{y}|\mathbf{a}_\tau, \mathbf{s})$ is intractable. However, it can be derived using the following approximations, similar to [5]. First, we follow Chung et al. [7] and employ an estimate of $\mathbf{a}_0$ at time τ as a sufficient statistic for $\mathbf{a}_\tau$ in the likelihood. This results in assuming $p(\mathbf{y}|\mathbf{a}_\tau, \mathbf{s}) \approx p(\mathbf{y}|\hat{\mathbf{a}}_0, \mathbf{s})$. The estimate $\hat{\mathbf{a}}_0$ is directly obtained from the unconditional score model via one-step denoising:

$$\hat{\mathbf{a}}_0 := \mathbf{a}_\tau - \sigma(\tau)^2 \mathbf{s}_\theta(\mathbf{a}_\tau, \tau, \gamma). \tag{6}$$

Furthermore, we model binaural reverberation by a convolution between the excitation signal $\mathbf{s}$ and our BRIR model $\mathbf{h}^{(\psi)}_{\hat{\mathbf{a}}_0}$, using the one-step denoising HRTF estimate $\hat{\mathbf{a}}_0$, and assuming $p(\mathbf{y}|\hat{\mathbf{a}}_0, \mathbf{s}) \approx p(\mathbf{y}|\mathbf{h}^{(\psi)}_{\hat{\mathbf{a}}_0} * \mathbf{s})$. Finally, we follow [19] and approximate the log-likelihood gradient using a L^2-distance in the magnitude-compressed STFT domain, between the measurement and our estimate:

$$\mathcal{C}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{M_\mathbf{y}} \| S_{\text{comp}}(\mathbf{y}) - S_{\text{comp}}(\hat{\mathbf{y}}) \|_2^2 , \tag{7}$$

where $S_{\text{comp}}(\mathbf{y}) := |\mathcal{F}_{\text{ST}}(\mathbf{y})|^{2/3} \exp j\angle\mathcal{F}_{\text{ST}}(\mathbf{y})$ is the magnitude-compressed spectrogram with $M_\mathbf{y}$ STFT frames. This compression accounts for the heavy-tailedness of speech distributions [20]. Note that we also use this function as the objective for optimizing the BRIR parameters ψ in (4). The log-likelihood gradient is finally obtained as:

$$\nabla_{\mathbf{a}_\tau} \log p(\mathbf{y}|\mathbf{a}_\tau, \mathbf{s}) \approx -\zeta(\tau) \nabla_{\mathbf{a}_\tau} \mathcal{C}(\mathbf{y}, \mathbf{h}^{(\psi)}_{\hat{\mathbf{a}}_0} * \mathbf{s}), \tag{8}$$

where $\zeta(\tau)$ adjusts the weight of the log-likelihood gradient during sampling, an is parameterized following [5], [8]. In conclusion, the posterior sampling procedure amounts to solving the following ODE:

$$\mathrm{d}\mathbf{a}_\tau = -\tau \left[\mathbf{s}_\theta(\mathbf{a}_\tau, \tau, \gamma) - \zeta(\tau) \nabla_{\mathbf{a}_\tau} \mathcal{C}(\mathbf{y}, \mathbf{h}^{(\psi)}_{\hat{\mathbf{a}}_0} * \mathbf{s}) \right] \mathrm{d}\tau, \tag{9}$$

In summary, the resulting algorithm alternates between optimizing BRIR parameters and estimating the HRTF, as illustrated in Fig. 1. At each step n of the discretized diffusion time axis, we perform $N_{\text{its.}}$ optimization iterations of the parameters ψ_n in our BRIR model (3), followed by a sampling step of the ODE (9) to update the HRTF estimate $\mathbf{a}_n$.

V. Experimental Setup

A. Experimental Data

HRTF data: We obtain the time-aligned features required to train the score model and evaluate our HRTF estimation method from the simulated HRTF sets of the HUTUBS database [21]. In practice, the pure delay component is estimated and removed from the channel of each HRTF data point. This yields 2×128-dimensional binaural time-aligned HRTF spectra after dropping the Nyquist bin (more details in [22], [23]). Out of the 95 HUTUBS subjects, 85 are used to train the score model. Two subjects are reserved for validation and six for testing. Repeated simulations "88" and "96" of HUTUBS subjects "1" and "22" are excluded from our splits.

TABLE I: *Instrumental results obtained on simulated 44.1-kHz BRIR data. Values indicate mean and standard deviation. Lower is better*

Method	LRE	LMD
Random	-3.43 ± 1.59	4.58 ± 0.69
Generic	-4.01 ± 1.75	4.77 ± 0.74
Proposed	**-9.20 ± 5.64**	**2.28 ± 1.11**
Nearest Neighbour	*-7.50 ± 1.46*	*3.76 ± 0.61*

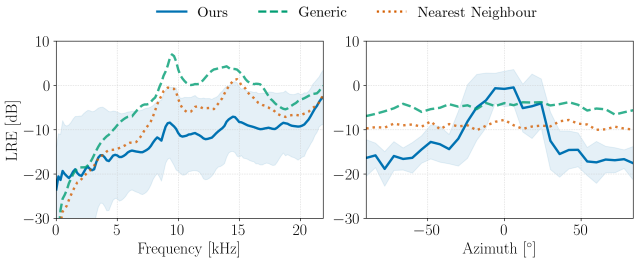
Estimation task data: We evaluate the performance of our HRTF estimation method using reverberant binaural speech observations generated by filtering utterances from VCTK's speakers "p226" and "p287" [24] with simulated BRIRs. The BRIRs are generated using a publicly available shoebox simulation software implementing the image-source method [25], [26], which we set to a reflection order of 20. We provide the software with our test HRTFs for binauralization of the room impulse response. Approximately 100 tasks were generated per test subject for a total of 599.

In each estimation task, the room's height is drawn uniformly in the $[2.5, 4]$ m range and the floor dimensions (width and length) in the $[7, 15]$ m range. The source and the subject's head locations is drawn uniformly within the volume of the room at a distance of at least 1.5 m from the walls and with a height in the $[1, 2]$ m range. The source is maintained at least 1 m away from the subject's head and its location is slightly adjusted so that it matches the closest DoA from the subject's HRTF set. Finally, the absorption coefficient of the simulation model is drawn within the $[0.05, 0.1]$ range. A similar validation set was generated for tuning the hyperparameters of our method using the HUTUBS subjects from our validation split. The VCTK recordings were down-sampled to 44.1 kHz prior to filtering so as to match the simulation's sample rate.

B. Implementation Details

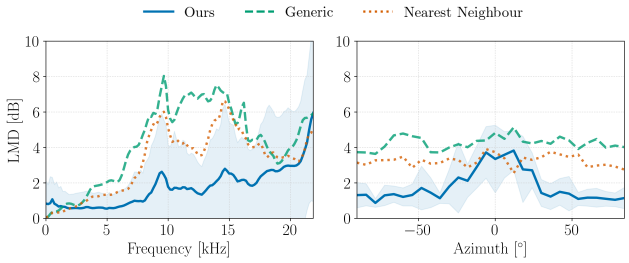
Diffusion Parameterization The score model is trained using diffusion times between $T_{\min} = 0.01$ and $T_{\max} = 10$. At inference time, we reduce the extremal times to $T_{\min} = 0.05$ and $T_{\max} = 8$ as they showed to be sufficient. The above levels are relative to normalized HRTF features, which are obtained by scaling the original features with respect to the mean and variance of our training set. We discretize the diffusion time axis into $N = 100$ steps for reverse diffusion using the logarithmic discretization in [13].

DNN Architecture: We parameterize the HRTF score model with a 1D-UNet comprising seven encoding-decoding stages, each with a resampling factor of 2 and comprising 32 hidden features. The model encodes the DoA γ and the diffusion time τ using random Fourier features embedding [27]. The above architecture results in a total of 752k parameters which we optimize using Adam with a learning rate of 5×10^{-4} and batch size of 32 for 110k steps. During training, we prevent the model from relying too heavily on DoA information by



(a) (b)

Fig. 2: *LRE as a function of (a) frequency and (b) azimuth.*



(a) (b)

Fig. 3: *LMD as a function of (a) frequency and (b) azimuth.*

injecting white noise ($\sigma = 0.05$) in the value of γ, and even completely dropping out γ with a probability of 30%. We track an exponential moving average of the DNN weights with a decay of 0.999.

BRIR Model: STFTs are computed using a Hann window of 23 ms and 75% overlap. We set the number of STFT frames of our reverberation operator to $M_{\mathbf{r}} = 200$, which corresponds to 120 ms. We decimate the frequency scale into $B = 40$ bands using a quasi-logarithmic spacing [6]. We optimize the BRIR parameters ψ with Adam [28], with a learning rate of 0.01, momentum parameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and $N_{\text{its.}} = 50$ optimization iterations per diffusion step. After each optimization step, we further clamp the weights $[w_b]_b$ between 0 and 40 dB, and the decays $[\alpha_b]_b$ between 0.01 and 40. This helps stabilize the optimization at early sampling stages. The parameters are initialized to $g = 0.15$, $t_{\text{left}} = 52$ samples, $w_b = 2$ and $\alpha_b = 0.1$ (across all frequency bands). Finally, t_{itd} is initialized according to the DoA.

C. Evaluation

We report the performance of our method in terms of logarithmic relative error (LRE)

$$\text{LRE}\left(\mathbf{a}_{c,f}, \hat{\mathbf{a}}_{c,f}\right) = 20 \log_{10} \left| \frac{\hat{\mathbf{a}}_{c,f} - \mathbf{a}_{c,f}}{\mathbf{a}_{c,f}} \right|, \tag{10}$$

and log-magnitude distance (LMD)

$$\text{LMD}\left(\mathbf{a}_{c,f}, \hat{\mathbf{a}}_{c,f}\right) = \left| 20 \log_{10} \left| \frac{\hat{\mathbf{a}}_{c,f}}{\mathbf{a}_{c,f}} \right| \right|. \tag{11}$$

We define three baselines for HRTF estimation: *Random* returns an HRTF filter drawn randomly from the training set at the specified DoA. *Generic* systematically returns HRTF filters from the HRTF subject forming a centroid of the training set, i.e. which minimizes the pairwise error to all the other HRTF subjects of the set as computed using the mean LRE across DoAs, frequencies and binaural channels. *Nearest Neighbour* is a quasi-oracle baseline: it selects, amongst the training set, the HRTF with matching DoA that yields the lowest mean LRE error to the true HRTF.

VI. Results and Discussion

The results of the objective metrics are reported in Table I. The proposed method outperforms the compared baselines, including the *Nearest Neighbour* oracle in both LRE and LMD metrics. Figures 2 and 3 illustrate the instrumental metrics as a function of frequency and azimuth. The results in Figures 2a and 3a reveal several key trends. First, the error increases with frequency, likely due to higher individual variability at higher bands. Notably, in the 5–8 kHz range, our method achieves a mean LRE that is at least 8 dB lower than the *Generic* HRTF baseline. Furthermore, in the higher frequency range (8–17 kHz), the proposed method improves over the *Nearest Neighbour* oracle baseline by at least 6 dB in LRE and 2 dB in LMD. This performance gain highlights that our method surpasses mere data retrieval capabilities, which we attribute to the modeling capacity of the score-based prior. At lower frequencies (0–1 kHz), the error is slightly higher than the *Generic* baseline and *Nearest Neighbour*. However, this occurs below the range in which lie the most salient monaural cues, in particular filtering from the pinna (>3 kHz) [29].

One area of concern is the increased LRE observed in Fig. 2b at the median plane, i.e. 0° azimuth. In terms of LMD, the proposed solution also suffers from lower performance at the median plane, but our method still outperforms the *Generic* HRTF and is on par with the *Nearest Neighbour* baseline in this worst azimuth case. This suggests that the phase estimation (only assessed in the LRE metric) seems to suffer more than the magnitude estimation in this region. This overall phenomenon may be due to inherent challenges in modeling HRTFs at this spatial position, but this warrants further investigation.

VII. Conclusion

This paper proposed a posterior sampling scheme for HRTF estimation using a diffusion-based prior and a parameteric reverberation model for approximating log-likelihood computation. Compared to previous approaches, the proposed method can use natural-sounding sources such as speech, requires only one measurement, and more importantly, it can operate in a variety of reverberant environments. At the exception of target directions in the median plane, our method largely outperforms the considered baselines, including an oracle recommender system. In relative terms, the time-aligned HRTF estimation error is particularly low in the high-frequency region, which we attribute to the expressivity of the diffusion prior.

REFERENCES

[1] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 99–102.

[2] J. Reijniers, B. Partoens, J. Steckel, and H. Peremans, "HRTF measurement by means of unsupervised head movements with respect to a single fixed speaker," *IEEE Access*, vol. 8, pp. 92 287–92 300, 2020.

[3] Z. Yang and R. R. Choudhury, "Personalizing head related transfer functions for earables," in *Proceedings of the ACM SIGCOMM Conference*, 2021, pp. 137–150.

[4] V. Jayaram, I. Kemelmacher-Shlizerman, and S. M. Seitz, "HRTF estimation in the wild," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–9.

[5] E. Moliner, J.-M. Lemercier, S. Welker, T. Gerkmann, and V. Välimäki, "BUDDy: Single-channel blind unsupervised dereverberation with diffusion models," in *Proc. IWAENC*, 2024.

[6] J.-M. Lemercier, E. Moliner, S. Welker, V. Välimäki, and T. Gerkmann, "Unsupervised blind joint dereverberation and room acoustics estimation with diffusion models," *arXiv preprint arXiv:2408.07472*, 2024.

[7] H. Chung *et al.*, "Diffusion posterior sampling for general noisy inverse problems," in *Proc. ICLR*, 2023.

[8] E. Moliner, J. Lehtinen, and V. Välimäki, "Solving audio inverse problems with a diffusion model," in *Proc. ICASSP*, 2023.

[9] K. Diepold, M. Durkovic, and F. Sagstetter, "HRTF measurements with recorded reference signal," in *Proceedings of the Audio Engineering Society 129th Convention*, 2010.

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, 2020.

[11] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. NeurIPS*, 2019.

[12] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*. Journal of the American Statistical Association, 2000, vol. 82.

[13] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Proc. NeurIPS*, 2022.

[14] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[15] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely, "Efficient representation and sparse sampling of head-related transfer functions using phase-correction based on ear alignment," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 12, pp. 2249–2262, 2019.

[16] E. A. P. Habets, *Speech Dereverberation Using Statistical Reverberation Models*. London: Springer, 2010, pp. 57–93.

[17] F. Menzer and C. Faller, "Investigations on modeling BRIR tails with filtered and coherence-matched noise," in *Proceedings of the 127th AES Convention*, 2009, p. 7852.

[18] J. Fagerström, N. Meyer-Kahlen, S. J. Schlecht, and V. Välimäki, "Binaural dark-velvet-noise reverberator," in *Proceedings of the International Conference on Digital Audio Effects*, 2024, pp. 246–253.

[19] E. Moliner, F. Elvander, and V. Välimäki, "Blind audio bandwidth extension: A diffusion-based zero-shot approach," *arXiv*, 2024.

[20] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," *Proc. IWAENC*, 2010.

[21] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses," *J. Audio Eng. Soc.*, vol. 67, no. 9, pp. 705–718, Sep. 2019.

[22] J. Nam, J. S. Abel, and J. O. Smith III, "A method for estimating interaural time difference for binaural synthesis," in *Proc. Audio Eng. Soc. Conv. 125*, Oct. 2008.

[23] E. Thuillier, C. T. Jin, and V. Välimäki, "HRTF interpolation using a spherical neural process meta-learner," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1790–1802, 2024.

[24] C. Valentini-Botinhao *et al.*, "Reverberant speech database for training speech dereverberation algorithms and TTS models," *University of Edinburgh*, 2016.

[25] R. Barumerli, D. Bianchi, M. Geronazzo, and F. Avanzini, "SofaMyroom: a fast and multiplatform "shoebox" room simulator for binaural room impulse response dataset generation," *arXiv preprint arXiv:2106.12992*, 2021.

[26] S. M. Schimmel, M. F. Muller, and N. Dillier, "A fast and accurate "shoebox" room acoustics simulator," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 241–244.

[27] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," in *Proc. NeurIPS*, 2020.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. ICLR*, 2015.

[29] C. Jin, A. Corderoy, S. Carlile, and A. van Schaik, "Contrasting monaural and interaural spectral cues for human sound localization," *J. Acoust. Soc. Am.*, vol. 115, no. 6, pp. 3124–3141, 2004.