Deep Learning-based Speaker Verification In Real Conditions

Sandipana Dowerah

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Deep Learning-based Speaker Identification In Real Conditions

THÈSE

présentée et soutenue publiquement le 30/05/2023

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Sandipana Dowerah

Composition du jury

Président :	Frédéric Sur	Professeur, Loria, Université de Lorraine.
Rapporteurs :	Sylvain Meignier	Professeur, LIUM, Le Mans Université.
	Sylvain Marchand	Professeur, IUT de La Rochelle.
Examinateurs :	Nancy Bertin	Chercheuse, Oracle.
Encadrants :	Romain Serizel	Maître de conférences, Université de Lorraine
	Denis Jouvet	Ancien Directeur de Recherches, Inria Nancy – Grand Est

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Résumé

Les applications telles que la vérification du locuteur sont devenues essentielles pour vérifier l'identité de l'utilisateur à partir de ses caractéristiques vocales pour des assistants personnels ou des services bancaires en ligne. Cependant, la vérification du locuteur avec une prise de son distante est constamment affectée par les bruits environnants qui peuvent considérablement déformer le signal vocal. De plus, les signaux vocaux sont réfléchis par divers objets dans la zone environnante, ce qui crée de la réverbération et dégrade encore plus la qualité du signal. Cette thèse explore les techniques de rehaussement de la parole à multicanal basées sur l'apprentissage profond pour améliorer les performances des systèmes de vérification de locuteur dans des conditions réelles. Le rehaussement de la parole multicanal vise à améliorer la qualité de la parole captée par plusieurs microphones. Elle est devenue cruciale pour de nombreux terminaux, qui sont flexibles et pratiques pour les applications vocales.

Trois approches novatrices sont proposées pour améliorer la robustesse au bruit du système de vérification de locuteur. Tout d'abord, nous intégrons une architecture de réseau neuronal profond avec des techniques de traitement du signal pour le rehaussement de la parole en tant que prétraitement d'un système de vérification de locuteur basé sur les x-vecteurs. Nous examinons l'importance d'effectuer aussi un prétraitement pendant la phase d'enrôlement du locuteur, ce qui a été largement négligé dans la littérature. L'évaluation expérimentale montre que le prétraitement les performances de vérification de locuteur si les fichiers d'enrôlement sont traités de manière similaire à ceux de test, et si le test et l'enregistrement se font dans des plages de signal à bruit similaires. Nous proposons ensuite de mettre en œuvre des modèles de diffusion probabilistes basés sur des scores pour le rehaussement de parole multicanal en tant que front-end d'un système ECAPA-TDNN de vérification de locuteur. Nous mettons particulièrement l'accent sur les techniques de rehaussement de parole multicanal. Nous utilisons des approches de diffusion probabilistes pour calculer soit des masques temps-fréquence, soit des filtres multicanaux. Comme l'entraînement séparé du module de rehaussement de la parole introduit souvent des artefacts et des distorsions, cela entraine une inadéquation pour la vérification du locuteur. Nous proposons une optimisation conjointe pour pallier à ce problème. Nous avons étendu les approches mentionnées ci-dessus en optimisant conjointement les modèles de rehaussement de la parole et de vérification de locuteur avec ou sans prise en compte d'un critère de distillation de connaissances. Ce critère de distillation de connaissances minimise la distance entre les plongements de locuteur obtenus à partir du système proposé et ceux obtenus à partir de signaux de parole propres (non bruités), améliorant ainsi les performances du système de vérification de locuteur dans différentes conditions de bruit.

Mots-clés: rehaussement de parole multicanal, vérification de locuteur avec prise de son distante, apprentissage profond

Abstract

Smart applications like speaker verification have become essential in verifying the user's identity for availing of personal assistants or online banking services based on the user's voice characteristics. However, far-field or distant speaker verification is constantly affected by surrounding noises which can severely distort the speech signal. Moreover, speech signals propagating in long-range get reflected by various objects in the surrounding area, which creates reverberation and further degrades the signal quality. This PhD thesis explores deep learning-based multichannel speech enhancement techniques to improve the performance of speaker verification systems in real conditions. Multichannel speech enhancement aims to enhance distorted speech using multiple microphones. It has become crucial to many smart devices, which are flexible and convenient for speech applications.

Three novel approaches are proposed to improve the robustness of speaker verification systems in noisy and reverberated conditions. Firstly, we integrate a deep neural network architecture with signal-processing techniques for speech enhancement as a pre-processing to an x-vector-based speaker verification system. We examine the importance of also using such pre-processing during the enrollment phase, which has been largely overlooked in the literature. Experimental evaluation shows that pre-processing improves speaker verification performance if the enrollment files are processed similarly to the test data and if the test and enrollment occur within similar signal-to-noise ranges. We then propose to implement novel score-based diffusion probabilistic models for multichannel speech enhancement as a front-end to an ECAPA-TDNN speaker verification system. Particular emphasis is put on multi-channel speech enhancement techniques. We compute the time-frequency masks and multichannel filters using diffusion probabilistic models. As individual training of the speech enhancement module often introduces certain artefacts and distortions, leading to mismatch problems. We propose joint optimization of both modules as it helps in retaining the information. We expanded the aforementioned approaches by jointly optimizing speech enhancement and speaker verification with and without knowledge distillation loss. The knowledge distillation loss minimizes the distance between the speaker embeddings obtained from the proposed system and those obtained clean speech signals, further improving the performance of the speaker verification system on different noise conditions.

Keywords: multi-channel speech enhancement, far-field speaker verification, deep learning

Acknowledgments

Acknowledgements play a vital role in every thesis, and I want to express my gratitude to those who have supported me throughout my PhD journey.

Firstly, I would like to express my sincere appreciation to my supervisors, Denis Jouvet and Romain Serizel, for their invaluable guidance, encouragement, and support throughout my three-and-a-half-year thesis work. Their expertise has been instrumental in shaping this thesis. Their constant motivation, constructive feedback, and understanding have helped me overcome challenges and doubts while working on this project. I am grateful for their patience and mentorship, which will stay with me for the rest of my career. I would like to thank them for providing their support throughout the thesis work as well as for other activities such as internship supervision and attending conferences and summer schools.

I also want to thank my friends and colleagues, Ajinkya Kulkarni and Ioannis Douros, for their important advice and constructive feedback throughout this journey. I would like to extend my gratitude to all the permanent members of the Multispeech team. I would like to express my love to my colleagues who became friends for life, especially Marina and Vinicius, for always being there for me, Cannette, Tulika, Sewade, Louis Delebecque, Arash, Pierre, Ashwin, Francesca, Marie-Anne, Ali, and Lou, for all the beautiful memories we shared in Nancy. I would like to thank Nicholas Furnon for his assistance during the beginning of my thesis work. I would also like to acknowledge the project partners from LIA, Avignon Driss Matrouf and Mohammad Mohammadamini, for all the discussions and help they have provided for carrying out this research.

I would like to express my gratitude to my CSI committee member, Frédéric Sur, for providing valuable advice and suggestions. Our discussions on the research topic, as well as other activities, have been tremendously helpful throughout this journey.

I would like to express sincere gratitude towards all the administrative staff members for their support in the documentation and administrative work, which includes Helene Cavallini, Annick Jacquot, Delphine Hubert, Anne-Marie Messaoudi, Emmanuelle Deschamps, Souad Boutaguermouchet, and Sabrina Ferry.

I would also like to acknowledge my parents, Aroti Gohain and Rabindranath Dowerah, and my sisters, Uddipana and Rituporna, for their continuous support and encouragement in pursuing my studies in Europe.

Finally, this thesis was funded by the National Research Agency as part of the Robovox project (ANR-18-CE33-0014). Some of the experiments presented in this manuscript were performed on Grid5000, a server supported by a scientific interest group of Inria, CNRS, RENATER, and other universities and organizations (see https://www.grid5000). Other experiments were carried out on the EXPLOR server at the University of Lorraine.

Dedicated to maa and deta...

Contents

List of Figures

List of Tables

List of Abbreviations

BLSTM ..... Bi-directional Long-Short-Term Memory

CNN ..... Convolutional Neural Network

CRNN ..... Convolutional Recurrent Neural Network

C.I ..... Confidence Interval

DNN ..... Deep Neural Network

DPM ..... Diffusion Probabilistic Model

EER ..... Equal Error Rate

FFT ..... Fast Fourier Transform

FIR ..... Finite Impulse response

GeLU ..... Gaussian Error Linear Unit

GRU ..... Gated Recurrent Unit

GMM ..... Gaussian Mixture Model

GEV ..... Generalized Eigenvalue

GEV-BAN ... Generalized Eigenvalue Blind Analytic Normalization

GLN ..... Global Layer Normalization

GSC ..... Generalized Sidelobes Canceller

HMM ..... Hidden Markov Model

KD ..... Knowledge Distillation

LCMV ..... Linearly Constrained Minimum Variance

LDA ..... Linear Discriminant Analysis

LSTM ..... Long-Short-Term Memory

MMSE ..... Minimum Mean Square Error

MFCC ..... Mel Frequency Cepstral Coefficient

MWF ..... Multi-channel Wiener Filter

MVDR ..... Minimum Variance Distortionless Response

NCC ..... Normalized Cross-Correlation

PSD ..... Power Spectral Density

PReLU ..... Parametric Rectified Linear Unit

PLDA ..... Probabilistic Linear Discriminant Analysis

ReLU ..... Rectified Linear Unit

RNN ..... Recurrent Neural Network

RIR ..... Room Impulse Response

SV ..... Speaker Verification

SNR ..... Signal-to-Noise Ratio

SAR ..... Signal-to-Artefact Ratio

SDR ..... Signal-to-Distortion Ratio

SIR ..... Signal-to-Interference

STFT ..... Short-time Fourier Transform

SDE ..... Stochastic Differential Equation

SDW-MWF . Speech Distortion Weighted Multi-channel Wiener Filter

SI-SNR ..... Scale-invariant source-to-noise Ratio

TF ..... Time-frequency

TDNN ..... Time Delay Neural Network

TCN ..... Temporal Convolutional Network

WPE ..... Weighted Prediction Error

Chapter 1

Introduction

Contents

1.1 Motivation

Speech is one of the most effective and widely used means of communication between humans. Speech not only conveys messages like instructions but also carries information at different levels, such as instructions, speaker identities, gender of the speaker, age of the speaker, and express emotions. Over the years, with the evolution of computers, electronic media, etc., technological advancement has changed human communication. Although unrealistic at one time, technology has enabled machines or robots today to converse with their creators by endowing them with the power of instructing them to carry out various tasks. The usefulness of speech has led to extensive technological research and the development of various speech-based applications. The rapid rise of portable or hands-free devices, such as smartphones, tablets, smart televisions, etc., has expanded the market for speech communication interfaces. Speech technologies have advanced to personalized digital assistants in almost all mobile devices, and the development of smart speakers like voice bots or virtual assistants has paved a new way for human-machine interaction. This human-computer interface requires speech-based tools like speech recognition or text-to-speech synthesis. With the increasing demand for such speech technologies, one important aspect is security. For example, robust speaker verification and anti-spoofing technologies are used to authenticate the user's identity.

As we live in a noisy world, our voice communication is constantly affected by surrounding noises which severely distorts the speech signal. This is one of the key challenges for speech technologies. For instance, while talking to our colleagues online or via cell phone, the heating, ventilation and air conditioning systems can degrade the audio quality. Also, while working from home, we observed that voices from other parts of the house might get in the middle of the conversation, interfering with the speech from other

speakers, distant music signals, etc. Moreover, even in a silent room, the propagation of the speech signal gets affected by the environment's acoustic properties that create reverberation. Such phenomena degrade speech quality and intelligibility, which further degrades the performance of speech-based applications like speaker verification. Although humans are good at differentiating communication interference, machines are not smart yet.

To facilitate the adverse effects of acoustic noise and room reverberation, speech technologies require strong speech processing algorithms for improving speech quality and intelligibility in real noisy-reverberant environments. Speech enhancement can be used as a pre-processing to speech-based applications. The goal of speech enhancement is to improve the perceptual quality of speech by estimating clean speech signals impacted by adverse effects of noise and room reverberation. Therefore, speech enhancement is essential for many real-world speech-based technologies for enhancing distorted speech signals. Although for many years, research in different methods and algorithms has been going on for speech enhancement for speaker verification, it remains a challenging task in far-field scenarios. Moreover, few studies focus on multichannel speech enhancement for far-field speaker verification.

1.2 Thesis objective

During periods of inactivity, using an autonomous mobile robot to monitor human presence in industrial premises is a cost-effective solution. The robot moves around the premises and analyzes the activity there. The robot is responsible for verifying the identity when a person is detected. In case of difficulty, the robot then contacts a human operator. One of the major objectives of this project is to take into account as realistically as possible the real conditions of use of the robot. Speaker verification in the context of a mobile security robot faces several challenges related to the remote verification of a person's identity in real conditions, which can reduce performance drastically, such as:

i Ambient noise: when the robot operates in a noisy environment, such as in the presence of background sounds or environmental noise, such as traffic sounds or other sources of interference, speaker verification accuracy can be compromised. This is because the noise can interfere with the quality of the audio signal, making it difficult for the system to extract and identify the unique characteristics of a person's voice.

ii Reverberation: the persistence of sound waves within the robot's environment after the original sound source has stopped emitting sound. This can create an acoustic effect that can impact the quality and intelligibility of the sound captured or produced by the robot's microphones and speakers. Reverberation can be particularly problematic in environments with hard, reflective surfaces, as the sound waves bounce around and create a complex pattern of echoes that can interfere with speaker verification performance.

iii Impact of short utterances in speaker verification. Short utterances can present a challenge for robot-based speaker verification systems. Since short utterances contain fewer speech sounds and less vocal variation, they may not provide enough distinct vocal features for accurate verification of speakers.

iv A mismatch between enrollment and test can occur when significant differences exist between the characteristics of the speech used during enrollment and the speech being tested. This can lead to errors in speaker verification, as the system may not recognize the speaker's voice due to differences in speech quality, accent, or other vocal characteristics.

Despite efforts to remove acoustic challenges, speaker verification performance degrades in far-field scenarios. This is mainly due to far-field audio signals being captured from a distance, which can decrease signal quality due to factors such as attenuation and directional effects. This can result in a degraded audio signal that can affect speaker verification accuracy. Additionally, the quality of the microphones used to capture far-field audio can also impact speaker verification performance. Low-quality microphones may introduce additional noise or distortion into the audio signal, which can reduce the accuracy of the system.

This thesis aims to propose and develop multichannel speech enhancement techniques as a front-end to remove ambient noise and reverberation for far-field speaker verification. The proposed solutions are based on signal processing and the exploitation of deep neural networks. Speech enhancement techniques are required for smart applications like speaker verification to improve the distorted quality of the signal by suppressing the background noise and room reverberation. Modern smart devices are equipped with microphone arrays exploiting the information acquired by the embedded multichannel in the devices. On the other hand, devices equipped with speech-based applications must ensure processing with computational efficiency. While deep neural networks are black boxes that require a lot of data and parameters and sometimes lack generalization, conventional signal processing techniques are constrained by the assumptions made about the signal statistics, which are often inaccurate and unrealistic. The algorithms proposed in the framework of this thesis are designed to integrate statistical signal processing and deep neural network techniques. As a result, we can take advantage of multichannel signal processing to design effective speech enhancement techniques that are high-performing and distortion-free. This can improve robustness while enabling the processing of speech signals in difficult non-stationary noisy situations.

1.3 Thesis contribution

We focus on developing various speech enhancement techniques as a pre-processing to a far-field speaker verification system. This thesis work mainly investigates three major frameworks for developing a robust noise-aware system: a benchmark of multichannel speech enhancement for far-field speaker verification, a score-based generative approach for multi-channel speech enhancement for far-field speaker verification, and the joint optimization of speech enhancement and speaker verification in a single framework. The main contributions of the thesis work are as follows:

1. We propose integrating filtering based on deep neural networks (DNN) or combining DNN and signal processing approaches. The DNN-based approach implements a neural beamforming technique for speech enhancement. The second approach integrates the neural beamforming technique with a signal processing-based classical multi-channel Wiener filter and a dereverberation technique. We integrated FaSNet, a deep neural network-based beamforming approach. We combine FaSNet

with Rank-1 multichannel Wiener filter and a weighted prediction error. We integrate this approach as a pre-processing pipeline to conduct various experiments for speaker verification in adverse acoustic conditions where noise and room reverberation distorts the target speech signal.

2. In addition to the aforementioned contributions, we conduct several studies concerning the impact of speech enhancement and speaker verification approaches in acoustic scenarios using various SNRs, the importance of enrollment processing, robustness to low SNR scenarios and generalization to unseen real recorded data. Additionally, we investigated the influence of quality (source-to-distortion and source-to-interference ratio) of the enhanced signals, which could help fine-tune the front end of a speaker verification system. We also study the impact of short utterance lengths on speaker verification performance.

3. We propose implementing a novel deep generative model for multichannel speech enhancement as a front-end for the speaker verification system in a far-field noisy-reverberant scenario. Our approach consists of score-based diffusion probabilistic models implemented for computing the masks or the multichannel filter. The score-based decoder learns to generate clean speech by gradually transforming the noisy multichannel signal using the Mel spectrogram or STFT. We quantify the relative importance of these main development techniques using speech enhancement and speaker verification objective metrics. We show that the final solution outperforms the popular state-of-the-art pre-processing techniques in all considered distortions.

4. To further improve the noise robustness of speaker verification, we propose a joint optimization approach. The individual training of the speech enhancement and the speaker verification system often leads to a mismatched problem. This problem is caused mainly by the speech enhancement module distorting some useful features, negatively impacting the back-end speaker verification model. However, joint training can retain the necessary features the speaker verification model requires, improving its performance. We jointly train speech enhancement and speaker verification in a single framework. The proposed joint optimization is a two-stage approach. We individually train the speech enhancement and speaker verification modules in the first stage. In the second stage, we jointly optimize both modules with and without the similarity-preserving knowledge distillation loss to better filter out the noise.

1.4 List of publications

The work carried out during this thesis led to the following publications:

- **Sandipana Dowerah**, Romain Serizel, Denis Jouvet, M.Mohammadamini, Driss Matrouf, How to Leverage DNN-based speech enhancement for multi-channel speaker verification?, In *International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI')*, 2022, Corfu, Greece.

- **Sandipana Dowerah**, Romain Serizel, Denis Jouvet, M.Mohammadamini, Driss Matrouf, Joint optimization of diffusion probabilistic-based multichannel speech enhancement with far-field speaker verification, *In IEEE Spoken Language Technology Workshop (SLT)*, 2022, Doha, Qatar.

- **Sandipana Dowerah**, Ajinkya Kulkarni, Romain Serizel, Denis Jouvet, Self-supervised learning with Diffusion based multichannel speech enhancement for speaker verification under noisy conditions, Submitted to *Interspeech*, 2023.

Other work done in collaboration with LIA, Avignon as part of the Robovox project are as follows:

- Mohammad Mohammadamini, Driss Matrouf, Jean-Francois Bonastre, Romain Serizel, **Sandipana Dowerah**, Denis Jouvet, Compensate multiple distortions for speaker recognition systems, In *European Signal Processing Conference (EUSIPCO)*, 2021, Dublin, Ireland.

- Mohammad Mohammadamini, Driss Matrouf, Jean-François Bonatsre, **Sandipana Dowerah**, Romain Serizel, Denis Jouvet, Le comportement des systèmes de reconnaissance du locuteur de l'état de l'art face aux variabiliés acoustiques, In *Journées d'Études sur la Parole (JEP)*, 2022, l'île de Noirmoutier, France.

- Mohammad Mohammadamini, Driss Matrouf, Jean-François Bonatsre, **Sandipana Dowerah**, Romain Serizel, Denis Jouvet, Learning noise robust ResNet-based speaker embedding for speaker recognition, In *Odyssey: The Speaker and Language Recognition Workshop*, 2022, Beijing, China.

- Mohammad Mohammadamini, Driss Matrouf, Jean-François Bonatsre, **Sandipana Dowerah**, Romain Serizel, Denis Jouvet, A Comprehensive Exploration of Noise Robustness and Noise Compensation in ResNet and TDNN-based Speaker Recognition Systems, In *European Signal Processing Conference (EUSIPCO)*, 2022, Belgrade, Serbia.

- Mohammad Mohammadamini, Driss Matrouf, Jean-François Bonatsre, **Sandipana Dowerah**, Romain Serizel, Denis Jouvet, Barlow Twins self-supervised learning for robust speaker recognition, In *Interspeech*, 2022, Incheon, South Korea.

1.5 Organization of the thesis

This thesis comprises a total of seven chapters, including this introductory chapter. The theoretical foundations of this Thesis and a review of the state-of-the-art are described in Chapter 2. The dataset and evaluation metrics are described in Chapter 3. Chapters 4, 5, and 6 are devoted to describing our contributions to multichannel speech enhancement as pre-processing to far-field speaker verification. Each chapter develops one of the previously enumerated objectives of this Thesis. Finally, Chapter 7 concludes. More specifically:

- Chapter 2 gives a detailed review of the state-of-the-art speech enhancement and speaker verification techniques used in the scope of this thesis work. We introduced the analysis and processing of the noisy reverberated speech signal. Then, the single-channel speech enhancement is reviewed on the problem of noise estimation. Next, multichannel speech enhancement approaches based on beamforming algorithms and multichannel filtering and dereverberation techniques are explained. To conclude, we give an overview of the use of deep neural networks for speech enhancement, specifically generative models.

- Chapter 3 describes the evaluation metrics and the datasets used in this thesis. We describe the objective quality metrics to evaluate the proposed speech enhancement and speaker verification contributions. We overview different datasets widely used for speech enhancement and speaker verification. Afterwards, we describe the speech enhancement and speaker verification datasets used in the experimentation of this thesis work. This includes the simulated dataset we developed in the framework of this thesis work, which is designed to simulate realistic room environments with additive noise and reverberation from clean and dry speech segments. Designing such a dataset is necessary as speech enhancement approaches require ground-truth knowledge about the target speech and, to some extent, the degradation. And this information is not available in the speaker verification corpus for far-field speaker verification.

- Chapter 4 explains the pre-processing pipeline for speaker verification in adverse acoustic conditions where noise and room reverberation distorts the target speech signal. The chapter describes the benchmark speech enhancement as a multichannel pre-processing approach for a far-field speaker verification system. We consider either filter based on deep neural network (DNN) or combining DNN and signal processing approach. This chapter explains the pre-processing pipeline we developed using three techniques: neural beamforming, multi-channel Wiener filter and dereverberation. We first make a general description of our approach, followed by an explanation of the neural beamforming architecture. Then, we describe the multichannel Wiener filter framework and, finally, the dereverberation algorithm we implemented to alleviate the reverberation. We study the impact of both approaches in different noisy and reverberated acoustic scenarios using various SNRs for a multi-channel input signal. We also study the impact of data mismatch, robustness in low SNR scenarios, and generalization to unseen real recorded data. Additionally, we investigate the influence of quality (in terms of source-to-distortion and source-to-interference ratio) of the enhanced signals, which could be helpful in fine-tuning the front-end of a speaker verification system.

- Chapter 5 describes the proposed diffusion probabilistic models-based methodologies for multi-channel speech enhancement. We analyse the diffusion-based techniques we developed to address acoustic interference and signal distortions as front-end processing for far-field speaker verification. We implemented a score-based diffusion probabilistic model to learn the time reversal using a stochastic differential equation for sample generation and gradually diffusing the data distribution towards a particular noise distribution. We propose several pre-processing techniques using diffusion probabilistic models for speaker verification. We provide a detailed explanation of the diffusion process, followed by an explanation of each technique. We present the experimental evaluation of our techniques on both synthetic and real-recorded corpora. Additionally, we evaluate the diffusion-based techniques on a newly released speaker verification dataset on two different enrollment conditions. For the diffusion-based techniques also, we evaluate the quality of the enhanced signals using the source-to-distortion and source-to-interference ratios.

- Chapter 6 describes the proposed joint optimization of speech enhancement and speaker verification framework. We propose two approaches for joint optimization that involve noise reduction techniques with and without knowledge distillation

loss. First, we explain each joint-optimised approach and then the loss functions we have used in this work. The first joint-optimized approach combines FaSNet, a neural beamforming technique, with the speaker verification system. The second approach utilizes diffusion-based multi-channel speech enhancement techniques in combination with the speaker verification system. We propose to use similarity-preserving knowledge distillation loss, type of knowledge distillation (KD) loss with the FaSNet-based joint-optimized model and with GradSE, a diffusion-based speech enhancement approach. The similarity-preserving knowledge distillation loss encourages the network to produce similar activation for enhanced signals. We use this technique to minimize the distance between speaker embeddings obtained from the proposed system and those of clean speech signals. We provide a detailed analysis of the experimental procedure as well as the performance of each approach.

- Chapter 7 summarizes the thesis and provides future research directions regarding multi-channel speech enhancement for far-field speaker verification.

Chapter 2

State-of-the-art

Contents

2.1 Introduction

"*We thought it was wrong to ask a machine to emulate people. After all, if a machine has to move, it does it with wheels—not by walking. Rather than exhaustively studying how people listen to and understand speech, we wanted to find the natural way for the machine to do it.*" - Fred Jelinek (IBM's speech recognition innovator).

The idea of designing a machine to mimic human communication and understanding has fascinated engineers and researchers and has been a topic of discussion and experimentation for centuries. Over time, computational and technological evolution has made it possible to communicate with inanimate objects and complete tasks through them rather than only being a fantasy. From Bell Laboratories AUDREY in 1952, which could recognize spoken numbers and the first voice synthesizer Voice Operating Demonstrator (VODER), to today's smart virtual assistants, speech technologies have witnessed a dramatically amazing transformation.

The goal of speech technologies is to convert recorded audio into a sequence of words. This has paved the way for speech technologies to help people with disabilities, transcribe interviews, learn a new language or access a file via voice commands. In the last few years, speech technology has emerged as one of the preferred methods for remote authentication due to the advances in telecommunications and networking and its ease of integration into existing systems. Automatic speech recognition (ASR), keyword spotting (KWS), language recognition (LRE), gender identification (GID), and emotions detection are just a few examples of the many approaches that make up automatic speech processing, each of which typically focuses on a single application. This thesis focuses on speaker verification and the pre-processing approaches for mitigating the factors that degrade its performance in a far-field setting.

Speaker verification is one of the ever-growing speech technology domains that has recently gained popularity. But the evolution of speaker verification goes a long way back. The history of speaker verification can be traced back to the early days of telephony when it became necessary to authenticate users for secure communication. Lawrence R. Rabiner and Biing-Hwang Juang, the researchers at Bell Laboratory, developed the first speaker verification system and recognized that voice recognition could be used to verify the identity of a speaker and prevent unauthorized access. They used pattern recognition and statistical modelling techniques to create a unique voiceprint for each speaker, which could be compared to a database of previously recorded voiceprints to verify the speaker's identity [Rabiner and Juang, 1993a].

In the 1970s and 1980s, speaker verification technology continued to evolve with the introduction of new algorithms and techniques for voice analysis. One major breakthrough during this period was the development of the hidden Markov model (HMM), a mathematical tool that could be used to model speech patterns and recognize individual speakers. HMM-based speaker verification systems were more robust and accurate than earlier systems, and they became widely used in commercial applications.

During the 1990s and 2000s, speaker verification technology continued to improve, with the introduction of new features such as mel-frequency cepstral coefficients (MFCCs) and Gaussian mixture models (GMMs). These techniques enabled speaker verification systems to handle more complex speech patterns and to adapt to individual speakers over time.

Today, speaker verification is used in a wide range of applications, from secure access control systems to voice-based authentication for online banking and other financial transactions. The technology continues to evolve, with ongoing research focused on improving

accuracy, robustness, and usability, as well as addressing new challenges such as deep fake audio and other forms of audio spoofing.

2.2 Problem formulation and notation

We consider an acoustic scene where the sounds emitted by different sources are affected by noise and reverberation. Let $y(t, f)$ be the pressure variation measured by K microphone at time-frequency t, f. The signal propagates in the room by reverberating on the different walls before reaching the microphone.

Considering the mixture of clean speech and noise as recorded by K microphones. The short-time Fourier transform (STFT) domain of the mixture is $K \times 1$ vector $\mathbf{y}$:

$$\mathbf{y}(t, f) = \mathbf{s}(t, f) + \mathbf{n}(t, f) \tag{2.1}$$

with

$$\mathbf{s} = \mathbf{h_s}(t, f)a(t, f) \tag{2.2}$$
$$\mathbf{n} = \mathbf{h_n}(t, f)b(t, f) \tag{2.3}$$

where $y_{t,f}$ is the noisy speech. $\mathbf{s} = [s_1 \dots s_K]$ is the $K \times 1$ spatial image of clean speech and $\mathbf{n} = [n_1 \dots n_K]$ contain the image of noise signal recorded at K. $h(t, f)$ is the room impulse response. Both noise and speech are assumed to be emitted by point sources. $\mathbf{h_s}(t, f)$ and $\mathbf{h_n}(t, f)$ are the room impulse response (RIR) from the speech and noise sources to the microphone array.

Multi-channel speech enhancement is based on the use of filters called beamformers. The filtered signal results from a convolution between the filter w and the temporal signal y, which can be more simply calculated under the narrowband approximation [Kowalski et al., 2010], by a scalar product in the complex domain;

$$y_{filt} = w^H y \tag{2.4}$$
$$= \sum_{k=1}^{k} w_k^* y_k \tag{2.5}$$

where the exponent H denotes the Hermitian transpose (the transpose of the complex conjugate), and the exponent $*$ denotes the complex conjugate of a scalar. Since the variables w and y are complex, the filter w modifies the amplitude and phase of each channel in y before summing them. Different types of beamformers are described in section 2.5.2.

2.3 Features

Feature extraction from the speech signal is the technique of extracting task-specific information from the speech waveform, also known as parameterisation. The fundamental goal of the feature extraction procedure is to turn the clean representation of speech signal into feature vectors that may effectively describe the signal's properties. Depending on the task being pursued, several parameter types are chosen. In the case of speaker verification, the feature extraction part must provide parameters resistant to noise and

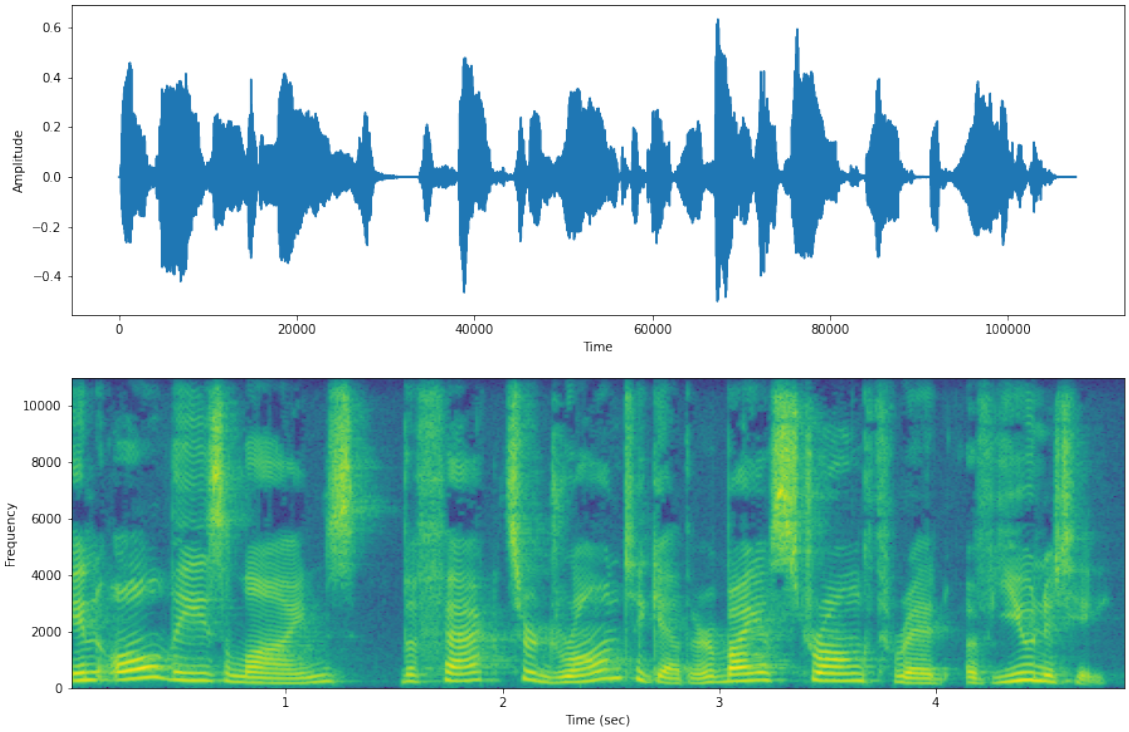

Figure 2.1: An example of speech waveform and associated spectrogram (magnitude of the STFT).

transmission effects. Figure 2.3 gives an overview of the graphical representation of the features for speaker verification.

We give a brief explanation of a few features in the following sections.

2.3.1 Time-frequency representation

Every sound is a vibration. Microphones digitally capture those vibrations, producing a waveform. Waveforms exhibit different structures at different time scales.

With the discrete Fourier transform (DFT), the discrete time signal is transformed into the frequency domain. Computations are usually carried on using the fast Fourier transform (FFT) algorithm [Oppenheim and Schafer, 2009]. The waveform is projected onto a basis of complex exponentials with linearly spaced frequencies. As DFT is invertible, all the information in the waveform is contained in its complex spectrum. Time-frequency representations were introduced to analyze the frequency content of the waveform over time. Contrary to time or frequency representations, time-frequency representations are sparsely distributed. As a result, mixtures of sounds tend to overlap less in the time-frequency domain, which makes separation easier. Figure 2.1 shows the STFT representation of a clean speech signal.

The short-time Fourier transform (STFT) is a widely known transformation commonly used for speech-processing tasks. The STFT of a clean signal can be defined as follows;

$$X(t,f) = \sum_{m=0}^{N_{s-1}} x(m+tl_s)w(m)e^{-J2\pi \frac{mf}{(N_s)}} \tag{2.6}$$

where m_s is the window function, N_s is the frame length in samples, l_s is the frameshift

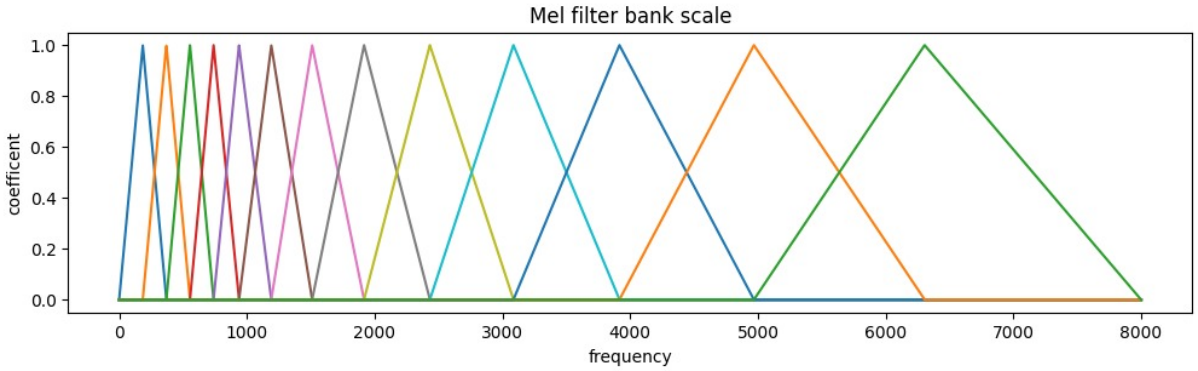

Figure 2.2: A graphical representation of Mel-scaled filter banks.

in samples, t and f are time and frequency indices which are multiplied by the window function. The DFT is then applied to each frame. An example of a clean speech discrete-time domain signal and its STFT representation is shown in Figure 2.1.

Time-frequency representation has been used for computing time-frequency masks as well. A time-frequency mask is a signal-processing tool to manipulate and enhance audio signals. In the masking-based methods, the objective is to generate a time-frequency mask which contains a weight for each time-frequency unit. Time-frequency masks are used to extract the target speech from noisy or reverberant signals in speech enhancement. The mask is estimated to correspond to the ratio of the target signal's magnitude spectrogram and the noisy or reverberant signal's magnitude spectrogram. The resulting mask is applied to the noisy or reverberant signal, effectively suppressing the noise or reverberation and amplifying the target speech.

They are often used as a pre-processing step before feature extraction and speaker verification. Researchers have proposed various algorithms for computing time-frequency masks, including Wiener filtering, spectral subtraction, and non-negative matrix factorization. Recently, deep learning-based approaches have gained popularity, where the time-frequency masks are estimated using convolutional neural networks or recurrent neural networks. These models have shown promising results in various noisy and reverberant environments.

Although spectrograms may occasionally be used directly as input features [He et al., 2016, Yu et al., 2019], the Mel spectrogram is usually preferred.

2.3.2 Mel filter banks

A Mel filter bank is a set of overlapping band-pass filters that are designed to mimic the non-linear frequency resolution of the human auditory system. Mel Filter Banks are used to provide a better resolution at low frequencies and less resolution at high frequencies. It captures the energy at each critical frequency band and gives approximates the spectrum shape. Mel filter banks are computed from FFT, then stacked together to form the Mel Spectrogram of the speech [Davis and Mermelstein, 1980, Zhang, 2017].

More specifically, to compute the Mel filter bank, a speech signal is applied to a pre-emphasis filter; this filter sectionalized speech signal into (overlapping) frames then a window function is applied to each frame. The pre-emphasis filter is useful to balance the frequency spectrum as high frequencies usually have smaller magnitudes compared to lower frequencies. After that, each frame is subjected to a Fourier transform, specifi-

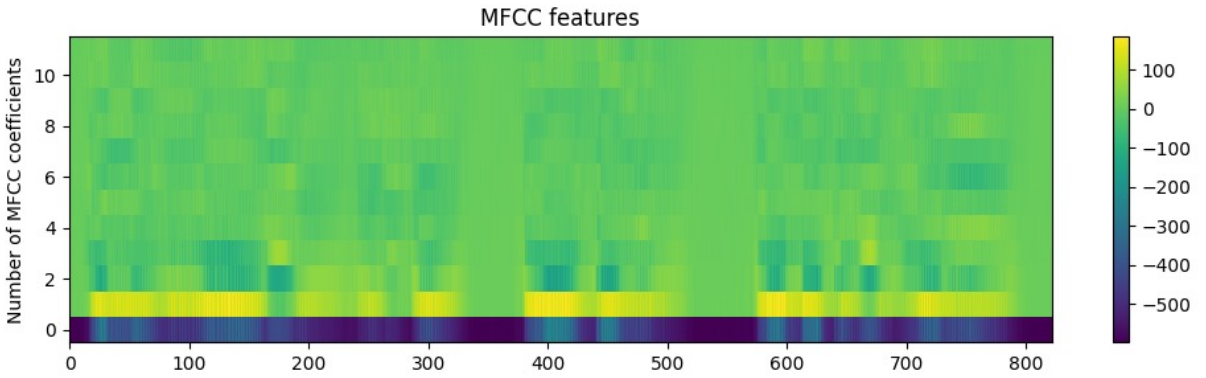

Figure 2.3: Example of MFCC features for a clean speech signal.

cally STFT preserving the magnitude spectrogram, which calculates the power spectrum. Hence, we get an approximation of the frequency spectrum of the speech signal by cascading adjacent frames. The final step in computing Mel filter banks is applying triangular filters on a Mel scale to extract frequency bands. Figure 2.2 is a graphical example of a Mel-scaled filter bank.

2.3.3 Mel-frequency cepstral coefficients

Mel-Frequency Cepstral Coefficients (MFCCs) were first introduced for speech recognition [Davis and Mermelstein, 1980]. The amplitude of the spectrum is initially acquired using the absolute value of the Fourier Transform. To better replicate the frequency resolution of the human ear by being more discriminative at lower frequencies and less discriminative at higher frequencies, the Mel-filters are then used to separate the spectrum into frequency bands [Stevens et al., 1937].

MFCCs are calculated by first applying a Mel filter bank to an audio signal to break it down into frequency bands. The energy of each band is computed as the weighted sum of squared values of the magnitude spectrum. A logarithm of the energies is then taken to compensate for the dynamic range of the values and to emulate the human perception of sound loudness. The resulting coefficients are transformed using the Discrete Cosine Transform (DCT) to obtain the cepstral coefficients, which represent the frequency content of the signal in a compact form. The inverse DCT is used to obtain a low-pass Fourier series representation of the frequency-warped log spectrum, where the fine structure corresponding to source information is filtered out. Finally, the dimensionality of the vector is reduced by the DCT to obtain the final vector of MFCC coefficients. Figure 2.3 is an example of MFCC coefficients for a clean speech signal. MFCCs have served as the acoustic feature of choice across speech applications.

2.3.4 Feature normalization

Speech signals and their features contain unique qualities that change depending on the recording environment (noise, reverberation, type of microphone). Convolutional noise shifts the mean values of MFCC coefficients, while additive noise reduces variance. To address this inter-session variability of characteristics, a straightforward mean and variance normalization has been proposed by Boll, 1979, and Openshaw and Masan, 1994. The entire utterance is normalized under the assumption that the undesired variability is

constant throughout the entire utterance. For the k^{th} frame in utterance d, the normalized i^{th} coefficient is computed in the following way;

$$\hat{c}_{d,i(k)} = \frac{c_{d,i(k)} - \mu_{d,i}}{\sigma_{d,i}} \tag{2.7}$$

where $\mu_{d,i}$ is the mean of the i^{th} coefficient, and $\sigma_{d,i}$ its standard deviation both estimated on utterance d.

To begin, we explain the basic notion of speaker verification, factors affecting speaker verification performance, and the various approaches adopted over the years to deal with degrading performance, especially in far-field or distant scenarios, while deploying in the real world.

2.4 Speaker verification

This section gives an overview of speaker verification over the years. As the main focus of this thesis is to develop multi-channel speech enhancement for speaker verification, we used the popular state-of-the-art speaker verification approaches for our experimentation without proposing any new approach. Therefore, we will look into the widely used approaches over time and gives a theoretical explanation of the three systems we have used in this thesis work.

Speaker verification is the process of verifying the authenticity of a person based on his/her voice characteristics. Each person's voice contains distinctive characteristics that computers try to take into account to automatically identify speakers' identity [Kinnunen and Li, 2010] automatically. According to Bai and Zhang, 2020, speaker verification is beneficial and has several uses. Speaker verification is typically used for security purposes, such as to verify that a speaker is indeed the person they claim to be when accessing a secure system. It can also be used in other applications, such as speech-based human-computer interaction and biometrics. For instance, speaker verification systems in the security field offer a step toward identity authorisation, increasing the security of an online payment system or a personal electronic device. Additionally, many downstream activities in speech technology can benefit from the speaker-related representation (also known as speaker embeddings) learned via a speaker verification system. For instance, a speaker-dependent automated speech recognition (ASR) system can exploit the speaker embeddings computed by a speaker verification system [Huang, 1991].

Theoretically, speaker verification has two distinctive steps: enrollment and verification. The enrollment phase in speaker verification is the process of collecting and storing a reference sample of a person's speech. During this phase, a speaker is asked to record a set of utterances which are then used to create a unique representation of the speaker's voice that can be used as a reference for future authentication attempts. This representation, known as the speaker model, is used to identify the speaker in future verification tasks. The enrollment phase aims to obtain a high-quality representation of the speaker's unique vocal characteristics, such as their speaking style, pronunciation, intonation, and rhythm. In a nutshell, enrollment is registering a speaker's voice biometrics and creating a speaker model.

Once the enrollment phase is complete, the verification phase begins. Verification uses the speaker model to determine whether a test speech signal matches the model associated with the claimed identity. During this step, a speaker is asked to pronounce some phrases or sentences used during enrollment. The system compares the new speech sample to the

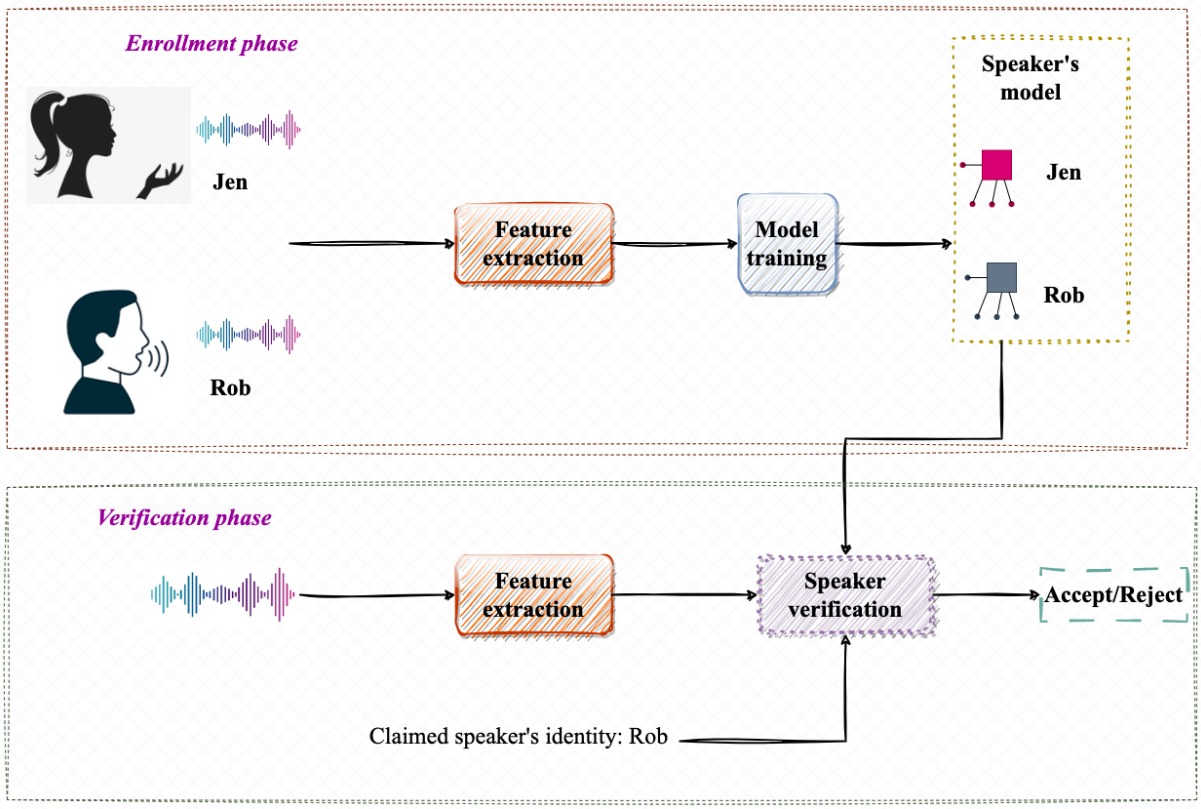

Figure 2.4: A generic speaker verification system consists of the following stages: Feature extraction extracts a set of speaker-discriminative features from the speech signal. Model training involves modelling the speaker-discriminative features to create a speaker model, represented as a set of speaker embeddings that capture the unique characteristics of the speaker's voice. Score computation computes a similarity score between the speaker embeddings obtained from the test speech signal and the speaker model. The score indicates the degree of similarity between the test speech signal and the enrolled speaker model. In the final stage, the system decides whether the test speech signal belongs to the enrolled speaker based on the computed similarity score. A threshold is typically used to determine whether the score is above or below a certain level, indicating a match or non-match.

reference samples stored during enrollment to determine whether the speaker is authentic. To do that, the system computes a similarity score, which indicates how closely the new speech sample matches the reference samples.

If the similarity score is above a predefined threshold, the system declares the speaker to be the same person who enrolled, and the verification is considered successful. If the similarity score is below the threshold, the verification is considered to have failed, and the speaker may be asked to repeat the process or provide additional authentication information. In short, if the score is sufficiently high, the claim is accepted, and if the score is low, the claim is rejected [Jin and Yoo, 2010, Campbell, 1997].

In simple words, speaker verification is a two-class problem that aims to determine if the speaker in the test utterance is the same as the speaker in the enrollment utterance. In modern speaker verification systems, acoustic features are extracted from each utterance and processed using VAD to remove any silence or non-speech segments. Then, these features are fed into a speaker embedding extraction module, which generates a fixed-

length vector representation of the speaker's unique characteristics, called an embedding. The enrollment and test embeddings are compared to determine whether they belong to the same speaker. This comparison can be made using various techniques such as cosine similarity, Euclidean distance, etc.

Speaker verification can further be distinguished as text-dependent or text-independent [Jin and Yoo, 2010]. In text-dependent speaker verification, the lexical content of the utterances is fixed or pre-defined, whereas in the case of text-independent speaker verification, there are no such conditions, and the speakers are free to speak as they want [Kinnunen and Li, 2010]. Unlike text-independent speaker verification, text-dependent speaker verification can rely on short utterances of a few seconds for authentication. A single speaker can utter several sentences, leading to multiple models per speaker. This is feasible because the speech content is controlled and limited to a brief sentence, needing only one utterance for the test. This allows the system to generate a more robust representation of the speaker's voice, which can improve the accuracy of the speaker verification process.

On the other hand, text-independent speaker verification systems typically require longer speech samples, usually of at least several seconds, to generate a robust representation of the speaker's voice. The exact length of the speech samples required for text-independent speaker verification can vary depending on the system and the specific application, but in general, speech samples of at least 30 seconds or longer are considered to be optimal for enrollment and test. We experimented with different utterance lengths and analysed the impact on speaker verification; the results are presented in Chapter 4. However, a reliable speaker model would be adaptable enough to verify the speaker using any utterance at the test time since there is no constraint on the speech content.

To review the foundational work in literature, we describe a few widely used approaches for speaker verification over the years.

Hidden Markov Models

Hidden Markov models (HMMs), a statistical approach, are eminent for identifying temporal patterns. HMMs are specifically used for their ability to characterise the stationary and temporal properties of a signal. HMMs have been widely used in speaker verification systems [Rabiner, 1989, Juang, 1991, Bahl, 1993, Gopinath, 2001, Shon, 2015, Zeinali et al., 2017], particularly in its text-dependent variation where the order of observations is more crucial, to describe phoneme transition probabilities [Heck and Genoud, 2001].

In speaker verification, HMMs are used to model the acoustic features of a speaker's speech. Each speaker is modelled as a separate HMM, with the acoustic features of their speech being used to estimate the parameters of the HMM. During verification, a test speech signal is compared to the HMMs of multiple speakers, and the speaker whose HMM results in the best match with the test signal is selected as the most likely speaker.

HMM assumes the speech signal as a parametric random process, and its parameters can be estimated precisely [Rabiner and Juang, 1993b]. HMMs model both the speech sounds and their temporal sequencing. HMMs model the speech feature vectors as a set of processes. The two stochastic processes carried out by HMMs are an observable process and a hidden Markov Chain (i.e., not observable directly, hence hidden process). According to the Markov property, the probability of adhering to any transition depends on the system's current state and is unaffected by previous observations. When building a model, a hidden Markov chain handles temporal fluctuations in the speech signal, while the observable process handles spectral variations.

HMMs have been effective in speaker verification, as they can capture the dynamic structure of speech signals and model the variability in the speech signals of different speakers. However, there are limitations to using HMMs in speaker verification, including the difficulty of accurately modelling the variability in speech signals and the limitations of the HMM framework in terms of modelling non-linear relationships in speech signals.

Deep Neural Networks

The era of using artificial intelligence (AI) for speaker recognition or verification dates long back after the idea of AI was proposed by numerous computer scientists in 1956s [Minsky, 1961]. But, due to limited computer hardware capabilities and the immaturity of AI algorithms, research on speaker recognition or verification did not produce promising results.

The use of Deep Neural Networks (DNNs) for speaker verification started to gain popularity around the mid-2010s due to their significant capacity, ability to learn complex patterns in the speech signal, strong feature extraction ability and flexible network architectures [Chen and Salman, 2011, Ghalehjegh and Rose, 2015, Variani et al., 2014, Snyder et al., 2018, Bai and Zhang, 2020]. The primary research on DNNs for speaker verification typically focused on replacing traditional GMMs with DNNs to extract speaker features. One of the key advantages of DNNs was that they could handle large amounts of data and model complex distributions, making them well-suited for speaker verification tasks.

A speaker verification system based on DNN extracts each speaker's deep properties and performs supervised learning in a neural network model when given labelled data. The typical approach for speaker verification using DNNs is to train the network on a large dataset of speech samples from multiple speakers. The network is trained to generate a speaker embedding, or a feature representation, for each speech sample. The embedding is then used to compare the similarity between speech samples from the same speaker and to differentiate speech samples from different speakers. The DNN-based architectures take in a sequence of acoustic feature frames, such as Mel-frequency cepstral coefficients (MFCCs), and learn to extract a speaker-discriminative representation. The network's final output is a fixed-length vector that summarizes the acoustic information for the entire speech sample.

Over time, many other papers have proposed different variations of DNNs for speaker verification, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and other variants of deep neural networks. DNNs have become a popular approach for speaker verification, and many state-of-the-art speaker verification systems now use DNNs to compute speaker embeddings.

2.4.1 Speaker representation

In speaker verification, dealing with different utterance lengths has always been a challenge. Finding a good representation of a recording independent of its length is one of the critical features of speaker verification systems. This is because it is important to compare the characteristics of the speech signal regardless of its duration. There are several methods for representing a speaker's voice for speaker verification. One of the earliest solutions for speaker representation to deal with different utterance lengths in speaker verification is to use a statistical model, such as Gaussian Mixture Models (GMMs) and likelihood comparison. In this approach, a GMM is trained to model the distribution of

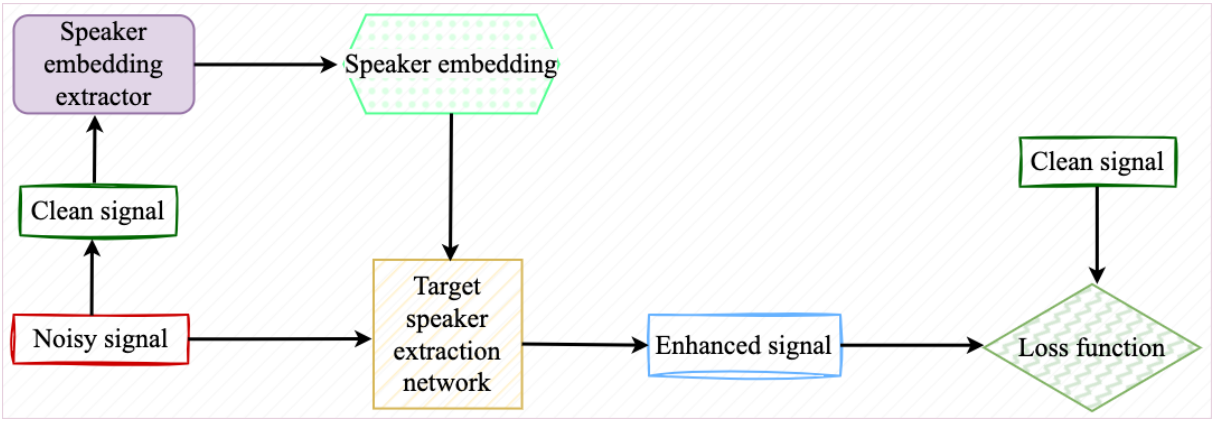

Figure 2.5: Graphical representation of the workflow of the speaker embedding extraction.

speech features for each speaker, and likelihood scores are then calculated for test utterances based on their fit to each speaker model. The use of GMMs and likelihood scores could be computationally intensive, particularly for large numbers of speakers or long test utterances.

In recent years, the development of fixed-length vector representations, or embeddings, which is a low-dimensional vector representation of a speaker's voice that is learned by a neural network, has become useful in various speech-processing applications. By providing a fixed-length representation of a recording, embeddings help to improve the accuracy and scalability of applications like speaker verification.

A detailed overview of the history and architectures of the classical approaches based on Gaussian Mixture Models and DNN-based approaches are given in the relevant sections below.

GMM-based approaches

Gaussian Mixture Model (GMM)-based approaches for speaker verification involve the speaker GMM, which is trained from one or several utterances of the given speaker(i.e., speaker training set). And then using the GMM to compute a likelihood score for a test utterance given a claimed speaker identity. This likelihood score is used to make a decision on whether the claimed identity is correct or not. GMM-based approaches were widely used in speaker verification and have been shown to be effective.

Gaussian mixture model-Universal background model

According to Reynolds and Rose, 1995, one of the earliest speaker recognition systems was built using the Gaussian mixture model (GMM). In GMM, the speaker is modelled using a combination of Gaussian probability density functions computed using the feature coefficients. GMM uses the expectation maximization (EM) algorithm [Muckenhirn et al., 2017], which is a generative modelling approach to iteratively bring the initially random values of the distributions closer to those of the observed data. When used for speaker recognition, this method which is frequently used to model multivariate data—produces a speaker-dependent GMM that can be assessed at various data points to see how well the speaker-dependent GMM resembles the data from an unknown speaker. One of the most basic iterations of the GMM-based technique involves creating a GMM for each known speaker.

Getting enough data in practice to train the GMM for each target speaker is challenging. Reynolds et al., 2000 suggested the Gaussian Mixture Model-Universal Background Model (GMM-UBM) as a solution to this issue. The GMM-UBM is a statistical method used in speaker verification that involves modelling the speaker-independent variability in the speech signal using a UBM and the speaker-specific variability using a GMM. The UBM is trained on a large speech corpus to model the background speech variability common to all speakers. During the enrollment phase, a speaker-specific GMM is trained for each speaker, using the UBM as a prior. This GMM is then used to represent the speaker-specific variability in the speech signal.

During the verification phase, the system compares the test speech signal to the speaker-specific GMM obtained during the enrollment phase. The system then calculates a likelihood ratio, which indicates how well the test speech signal fits the speaker-specific GMM relative to the UBM. If the likelihood ratio is above a predefined threshold, the speaker is considered the same person who enrolled, and the verification is successful. Traditional systems train the speaker GMM using the EM algorithm, and Maximum likelihood EM is used for UBM. The target speaker's GMM is then modified using the maximum a posteriori (MAP) method [Greig et al., 1989] by adjusting the parameters of the trained UBM rather than directly training the GMM for each speaker.

GMM-UBM is widely used in speaker verification systems, as it provides a simple and effective way to model the speaker-independent and speaker-specific variability in speech signals. The method is effective for both text-independent and text-dependent speaker verification. It can be combined with other speaker verification methods, such as Joint Factor Analysis (JFA), to improve the system's accuracy.

According to Markel et al., 1977, the effectiveness of a recognition system can be increased by creating and analyzing fixed-dimensional representations of each utterance. This made it possible to use various classifiers from different machine learning studies, such as the GMM supervector, which is created by concatenating the parameters of a GMM model to create a fixed-dimensional vector from a variable-duration utterance [Kuhn et al., 1998, Kenny et al., 2003].

Furthermore, conventional GMM-UBM systems are highly sensitive to variations in the utterances, whether from the channel or the speaker. This is because GMM-UBM systems model each speaker and each channel as separate GMMs, which can be sensitive to variations in the speech signals.

Supervectors

Supervectors are the first concept of representing variable-length utterances by a fixed-length vector. They are formed by concatenating the mean vectors of a GMM that has been trained on speaker-specific acoustic features, such as Mel frequency cepstral coefficients (MFCCs). Each mean vector corresponds to a mixture component in the GMM, and the resulting supervector has a dimensionality equal to the number of mixture components times the dimensionality of the acoustic feature vectors. The supervector representation provides a more robust and discriminative representation of the speaker's vocal characteristics.

According to Campbell et al., 2006, GMM supervectors can be successfully deployed for speaker verification using support vector machines (SVM) [Cortes and Vapnik, 1995], with the positive examples being the supervectors produced from the training utterances and the negative examples being a collection of fake utterances. However, the large

dimensionality of supervectors is a major issue, as it leads to reduce the performance of the back-end classification module [Campbell et al., 2006].

Joint factor analysis

Joint Factor Analysis (JFA), first introduced by Bourlard, 1998, is a projection of the supervectors in a reduced dimensional space and in improving the robustness of speaker recognition systems to variations in the speech signals [Kenny, 2005]. JFA is a statistical method for speaker verification that combines factor analysis and speaker modelling. The JFA technique has been widely studied and has been shown to be effective in speaker verification tasks. In particular, JFA has been used to improve the performance of speaker verification systems in noisy and challenging environments, such as background noise, channel distortions, and other environmental factors.

In JFA, a set of acoustic features, such as Mel-frequency cepstral coefficients (MFCCs), are extracted from speech signals and used to model the speaker-independent variability. The speaker-dependent variability is then modelled by combining the speaker-independent variability with a speaker-specific factor estimated based on the enrollment data.

During the verification phase, the system compares a test speech signal to the speaker-specific factors obtained during the enrollment phase. The system then calculates a similarity score, which indicates how closely the test speech signal matches the speaker-specific factors. If the similarity score is above a predefined threshold, the speaker is considered the same person who enrolled, and the verification is successful.

Dehak et al., 2011 used JFA to extract speaker features for speaker recognition. In their approach, JFA was used to combine the speaker and channel factors into a single space called the total variability space, as the authors realized that channel factors could contain speaker-dependent information. Dehak et al., 2011 stated that a speaker and session-dependent GMM supervector could be generated from the total variability space and from the hidden variables. The hidden variables in the total variability space, known as the total factors, were estimated by their posterior expectation and used as features for the next stage classifiers. These later became known as identity vectors, or i-vectors, in speaker recognition systems.

i-vectors

i-vectors introduced by Dehak et al., 2011 is a widely used feature representation type in speaker verification systems. An i-vector is a low-dimensional representation of a speaker's speech, which captures the speaker-specific information. The authors claim that the complete variability space and the hidden variables can create a speaker and session-dependent GMM supervector. Speaker or session variability is the variability exhibited by a given speaker from one recording session to another. Although not observable, the hidden variables, initially referred to as the total factors, can be estimated by their posterior expectation and used as features for the following stage classifiers.

The main aim of the i-vector paradigm is to use factor analysis to compress multiple sources of speech variability (such as speaker characteristics) into a fixed-length, low-dimensional representation. i-vector transforms the high-dimensional supervectors into a low-dimensional subspace by retaining the significant variability.

In speaker verification, i-vectors are derived from a UBM, which is trained on a large amount of speech data from multiple speakers. The UBM is used to model the spectral

patterns of speech, and the i-vectors are derived from the UBM by projecting the super-vector of a speaker's speech onto a low-dimensional subspace called the total variability space. The i-vector representation is speaker-specific, meaning that each speaker will have a unique i-vector.

A probabilistic linear discriminant analysis (PLDA) backend compensates for undesired channel characteristics in most approaches, providing a mechanism for classifying i-vector pairs to distinguish whether they belong to the same speaker or to different speakers. Although i-vectors have shown significantly impressive performance over unadapted models, they are quite disadvantageous in mismatched conditions as more enhancement information might be needed for pre-processing [Peddinti, 2015].

The i-vectors and PLDA have dominated the text-independent speaker verification in the past decade due to their superior performance, simplicity, and effectiveness. In this framework, a GMM-UBM [Reynolds et al., 2000] is used to collect sufficient statistics. Then, a feature extractor such as factor analysis [Dehak et al., 2011] is used to exact the low-dimensional identity vector as the compact representation of the utterance. The verification scores for each pair of utterances are then generated using a separately trained PLDA classifier. However, the performance of the system decreases for the short utterances in the enrollment/test [Zhang et al., 2018, Poddar et al., 2018].

In recent years, many researchers have proposed various modifications and improvements to the basic JFA technique. For example, some researchers have proposed using non-linear techniques, such as Deep Neural Networks (DNNs), to model the speaker and channel factors. Other researchers have proposed using Bayesian techniques to estimate the speaker and channel factors.

2.4.2 DNN-based approaches

DNN-based speaker verification systems have recently been developed, demonstrating competitive performance compared to conventional i-vector/PLDA systems. The DNN-based speaker verification systems typically use a neural network to extract a compact, speaker-discriminative representation of the speech signal. The neural network is learned to recognize (classify) the speakers of the training set. Once the training is done, when the NN is applied to some utterance, the information from the last hidden layer (to check whether it is the last hidden layer or the one before) is used as a fixed-length vector representation (the speaker embedding). This speaker embedding is then used as a feature representation to compare speech samples and decide whether they come from the same speaker or from different speakers.

The speaker embedding are used to compute speaker similarity between utterances [Li et al., 2017]. For example, if two speech samples are from the same speaker, their corresponding speaker embeddings should be close to each other in the feature space. On the other hand, if two speech samples are from different speakers, their corresponding speaker embeddings should be far apart in the feature space.

DNN architectures, including time-delay neural network (TDNN) [Snyder et al., 2018, Garcia-Romero et al., 2019], CNN [Bian et al., 2019, Zhang et al., 2018, Fang et al., 2020], and LSTM [Heigold et al., 2016, Wan et al., 2018] are used to extract frame-level features from utterances. The remaining part of the DNN corresponds to a pooling layer that merges the frame-level features of the utterance, followed by several feed-forward layers.

The DNN-based systems are trained using a classification loss, such as softmax loss [Nagrani et al., 2017] or angular softmax loss [Cai et al., 2018]. A few metric learning losses

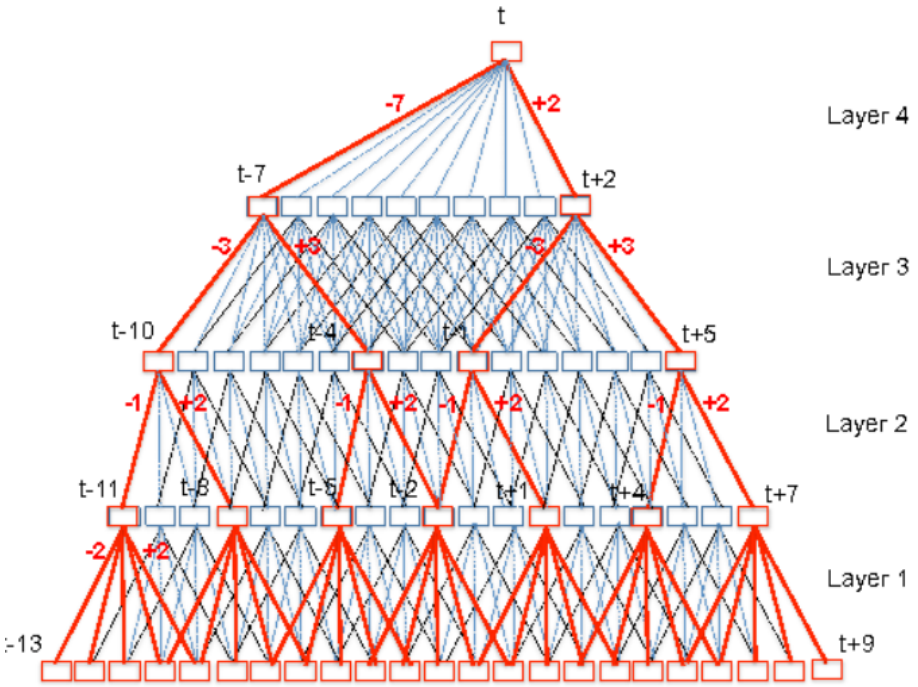

Figure 2.6: Graphical representation of the architecture of TDNN.

have also been exploited to train the entire speaker verification system in an end-to-end fashion, including triplet loss [Zhang et al., 2018, Li et al., 2017], generalized end-to-end (GE2E) loss [Wan et al., 2018], and cluster-range loss [Bian et al., 2019]. In addition, a lot of research on robust low-level features [Abdalmalak and Gallardo-Antolín, 2018, Al-Qaderi et al., 2021] and hybrid models [Shahin et al., 2021] has been done further for improving the effectiveness of conventional and DNN-based speaker verification systems.

Time delay neural network

Let's take a deeper look at the type of network that manages the context of input characteristics before we discuss the x-vector architecture in detail. Time Delay Neural Network (TDNN), first introduced by Waibel et al., 1989, is a multi-layer feed-forward deep neural network. It was later improved for better training efficiency by Peddinti et al., 2015, who used it as part of an acoustic model.

The motivation for using TDNN was to replace Recurrent Neural Networks (RNN) in the learning of long-term time dependencies. Because the learning mechanism is sequential, RNNs require more training time than feedforward networks. TDNN is used to learn long-term time dependencies from long contexts, and it achieves a similar learning speed to other feed-forward networks. Figure 2.6 illustrates the architecture of the TDNN network.

In one layer of the TDNN, each input frame is a column vector representing a single time step in the signal, representing the feature values. TDNNs learn to transform narrow context only in the first layer. The network uses a smaller matrix of weights (the kernel or filter), which slides over the signal and transforms it into an output using the convolution operation.

Deeper layers process hidden activations from a broader temporal context, hence the ability of the higher layers to learn broader temporal relationships. The layers have

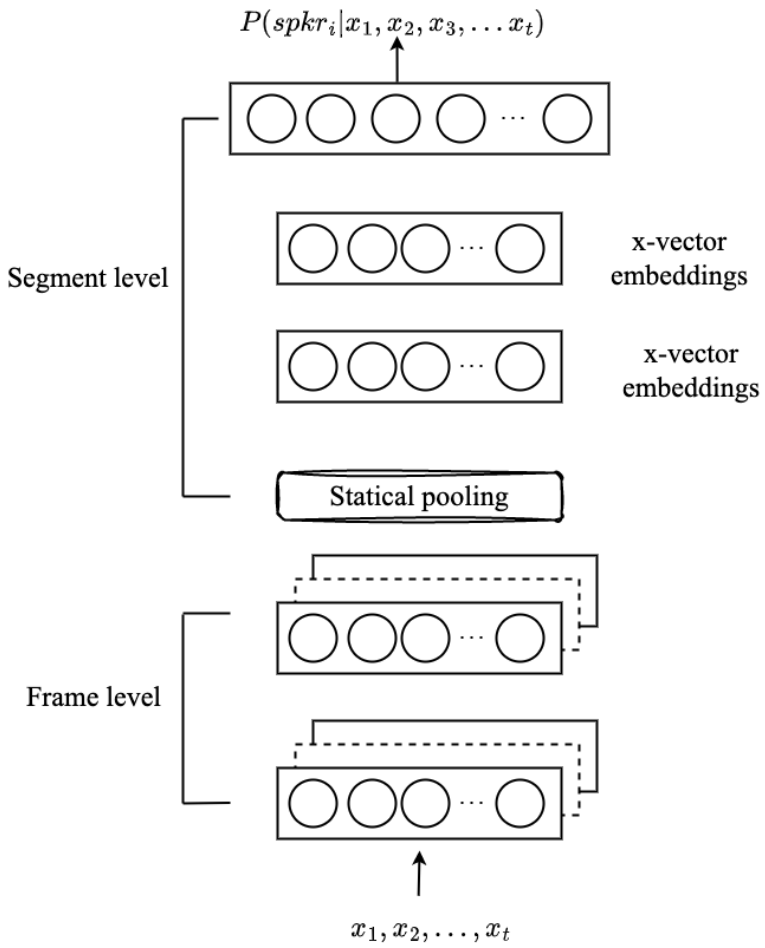

Figure 2.7: Graphical representation of x-vectors. The x-vector embeddings can be extracted after the statistical pooling network.

changing temporal resolution, which is higher with each deeper layer since each layer operates with a different width of context.

The bottom layers of the network are updated during back-propagation by a gradient accumulated throughout all the time steps of the input temporal context since the transformations in the TDNN architecture are connected across time steps. The TDNN is defined by the context of each layer it uses to calculate its activation.

X-vectors

X-vectors [Snyder et al., 2018] were introduced for extracting fixed-length feature vectors from recordings of any length. The backend of the speaker verification was trained separately to compare pairs of embeddings. The topology of the neural network is depicted in Figure 2.7. The network consists of three main parts: the frame level consists of layers operating on speech frames. The statistics pooling layer that aggregates over the frame-level representations to obtain a segment-level representation. And additional layers with the classifier. The additional layers operate at the segment level, and finally, a soft-max output layer provides the means for training the whole network via multi-class cross-entropy. Throughout the whole network, Rectified linear units are used as activation functions.

The original x-vector network has five frame-level TDNN layers. By assuming t as the current time step. At the input, frames are spliced together at $\{t-2, t-1, t, t+1, t+2\}$. The next two layers splice the previous layer's output together at times $\{t-2, t, t+2\}$ and $\{t-3, t, t+3\}$. The following two layers operate on the frame level but without any added temporal context. The size of the layers varies from $512-5136$ based on the used splicing context. Overall, the frame-level part of the network prepares frame-level representation from the time context of $\{t-8\}$ to $\{t+8\}$. The statistics pooling layer receives the output of the final frame-level layer as an input, aggregates input segments over time and computes their mean and standard deviation.

These segment-level statistics are concatenated and forwarded to two additional hidden layers with a dimension of 512. The two layers are used to compute the embedding (x-vector). The last layer consists of a softmax output. Segment-level embeddings can be extracted from any layer of the network after the statistics pooling layer.

The x-vectors can exploit large amounts of training data which makes it useful to augment or boost the data, termed data augmentation. Data augmentation increases the quantity of the data. The training samples are enhanced with noise and reverb using data augmentation, along with the original samples. Using the same front-end (feature extraction) or back-end (vector comparison) for both i-vector and x-vector systems enable system integration.

ECAPA-TDNN

The ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation) architecture was presented by Desplanques et al., 2020, which is a variant of x-vectors. The authors made architectural enhancements to the TDNN architecture and statistics pooling layer in the x-vector network. The temporal context of frame layers in the original x-vector system is limited to 15 frames. As it is beneficial to use a wider temporal context, Squeeze-and-Excitation blocks were introduced to represent channel inter-dependencies. The first component of a Squeeze-and-Excitation block is the squeeze operation which generates a descriptor for each channel. The squeeze operation consists of calculating the mean vector of the frame-level features across the time domain. The descriptors in the mean vector are then used in the excitation operation to calculate a weight for each channel.

Figure 2.8 illustrates the architecture of the ECAPA-TDNN network. The temporal attention mechanism has been extended to the channel dimension to enable the network to focus more on speaker characteristics that do not activate on identical or similar time instances. By scaling all the channels in accordance with the general characteristics of the recording, the Squeeze-and-Excitation block increases the temporal context of the frame layer. Moreover, as neural networks can learn hierarchical features by operating each layer at a distinct level of complexity, additional skip connections were introduced to propagate and aggregate the channels. Additionally, channel-dependent frame attention was incorporated to improve the statistics pooling layer. As a result, the network can concentrate on various subsets of frames while estimating each channel's statistics.

2.4.3 Scoring speaker similarity

Scoring is a critical component of speaker verification, as it is used to evaluate the similarity between the speech signal of a claimed speaker and the reference model of that speaker.

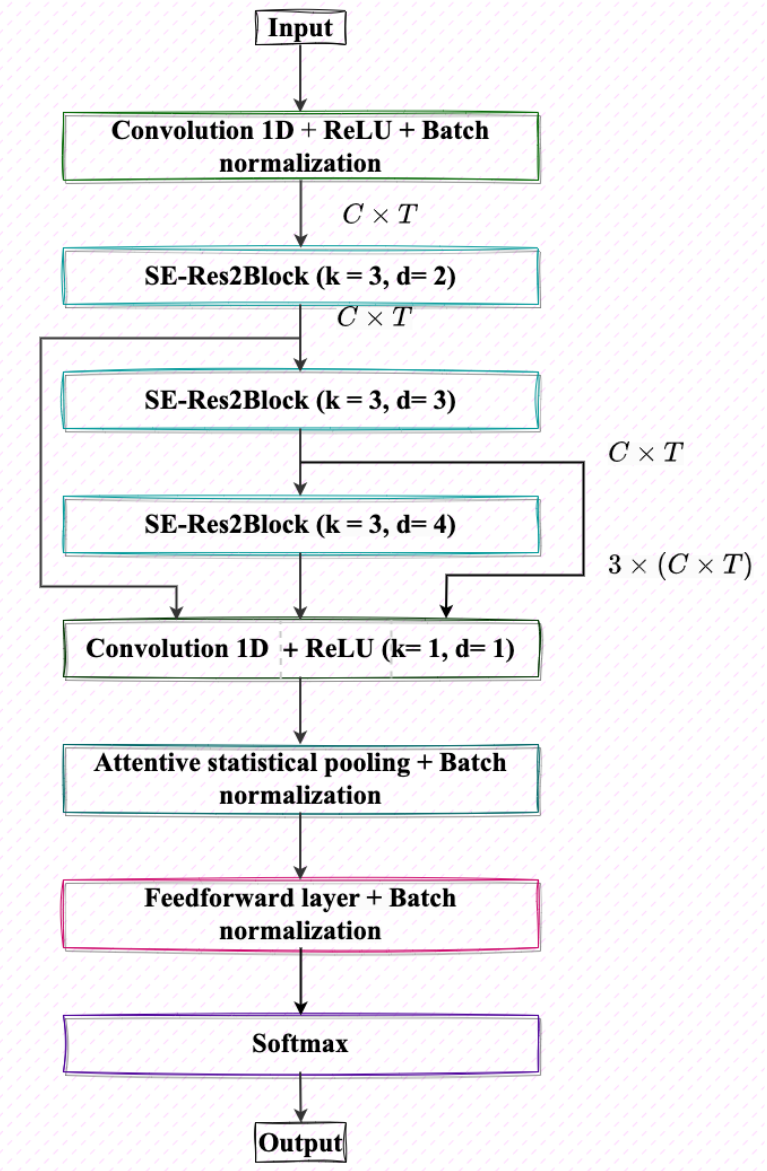

Figure 2.8: Network topology of the ECAPA-TDNN. k denotes for kernel size and d for dilation spacing of the Conv1D layers or SE-Res2Blocks. C and T correspond to the channel and temporal dimensions of the intermediate feature maps, respectively.

Speaker verification generates a match score by comparing the pre-enrolled speaker profile and incoming audio, indicating how similar the comparison results are, and utilising a threshold to determine whether to accept or reject the input as a match. Understanding the trade-off between false positive and false negative rates is crucial.

Two similarity measures are frequently used in speaker verification to determine whether a test observation is from the target speaker. These metrics are used by almost all novel DNN approaches, namely the cosine distance of vectors and probabilistic linear discriminant analysis (PLDA). This section explains the techniques we have used to compare two embeddings.

Cosine similarity score

One of the first use of cosine similarity for speaker verification was made by Dehak et al., 2011. They used an SVM classifier for i-vectors, using cosine similarity distance for two vectors as the kernel function for the SVM classifier.

In several verification tasks, cosine similarity scoring is used as it is a computationally effective method for such tasks. The decision score for speaker verification is calculated as the cosine of the angle between two embeddings, i.e., the enrollment ϕ_e and test embeddings ϕ_t. The angle represents the distance in the high-dimensional embedding space. The closer the angle is to 0 degrees, the more similar the embeddings are to each other and the higher the cosine similarity score. The cosine similarity score is computed as follows;

$$S(\phi_e, \phi_t) = \frac{< \phi_e, \phi_t >}{||\phi_e||||\phi_t||} \tag{2.8}$$

The cosine scoring method is advantageous as the scoring lies in low computational complexity. It should be noted that the scores produced in this manner are symmetric and unaffected by any potential vector swapping between the test and enrollment. Additionally, scoring is performed directly in the speaker embedding space.

Probabilistic linear discriminant analysis

Probabilistic linear discriminant analysis (PLDA) is a generative model, initially developed for face recognition [Prince and Elder, 2007], and is a probabilistic version of Linear Discriminant Analysis (LDA) that can handle more complexity in data. PLDA assumes that given data samples are generated from a distribution. PLDA has been widely used for recognition, verification, generating similarity scores for clustering, and class-specific feature extraction over time.

PLDA has become one of the popular techniques for comparing embeddings in verification jobs. PLDA aims to model speaker and channel-dependent variations in a lower-dimensional subspace. The i-vector/PLDA method is appealing and has advanced to state-of-the-art due to the reduced number of parameters compared to other approaches like joint factorization analysis. Moreover, PLDA can generate well-calibrated likelihood ratios without needing score normalization when training and evaluation data are taken from the same domain.

PLDA-based systems developed mainly for text-independent tasks, have been demonstrated to benefit from lexical information during training [Larcher et al., 2013]. However, they are inadequate for text-dependent tasks, especially when dealing with short utterances, due to their inability to model the temporal structure of the utterances [Larcher et al., 2013, Kenny et al., 2013]. PLDA can be used in new DNN approaches, where i-vectors are replaced with their deep learning alternatives (x-vectors).

The comparison score generated by PLDA is the log of the ratio between the likelihood that the embeddings were created by the same speaker and the likelihood that different speakers created them. The embeddings are assumed to belong to the same speaker if the comparison score is higher than the threshold; otherwise, we claim they belong to different speakers.

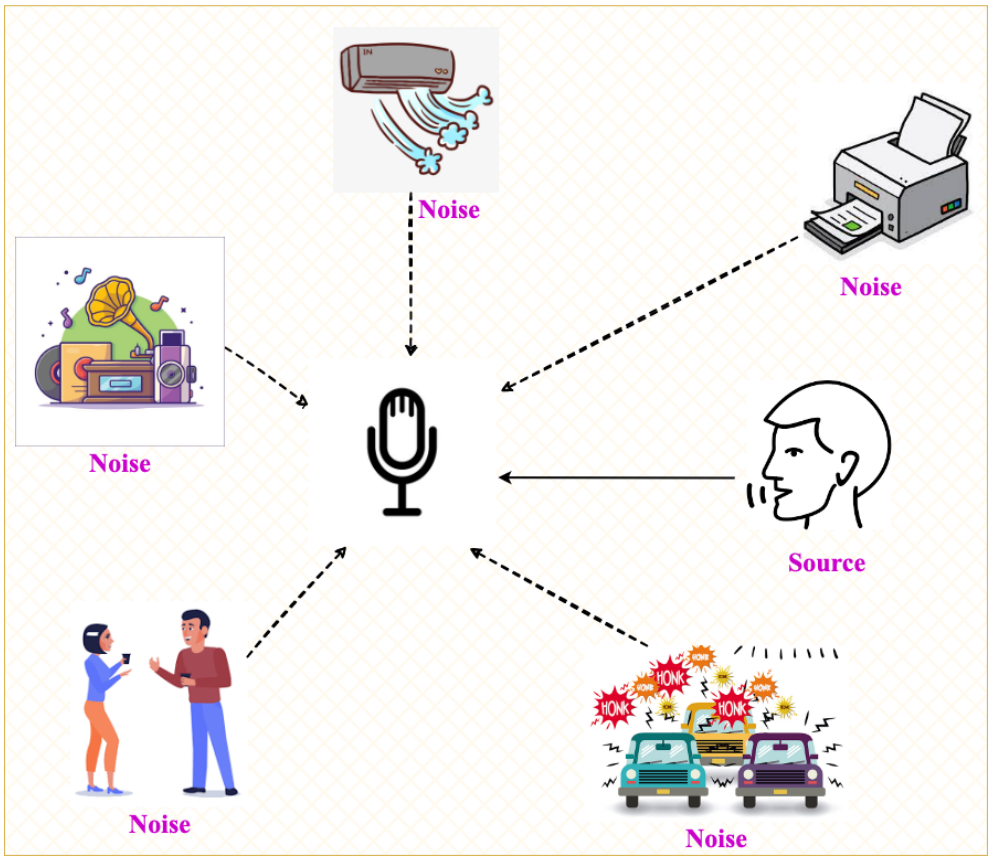

Figure 2.9: Graphical illustration of a noisy environment.

2.4.4 Factors affecting far-field SV

There are several factors that can affect the accuracy of speaker verification systems. It is not possible to provide a perfect recording environment free of background noise, reverberation, and other acoustic and environmental disturbances. Additionally, recording the speaker's voice using high-quality microphones might not always be feasible. This section overviews the signal degradation only in acoustic environments, among other factors. We can approximate an acoustic environment by two types of signal degradation: noise and reverberation. Usually, in literature, the term "noise" represents all types of signal degradation (including reverberation). In this work, we distinguish between noise and reverberation.

Noise

Several types of noise can affect speaker verification systems, as shown in Figure A.1. Background (or ambient) noise in real-world settings always distorts speech signals. This distortion can take many forms, including additive noise [Ming et al., 2007]. It can come from both external and internal sources. Here are some examples: environmental noise is any unwanted sound that is present in the environment in which the speaker verification system is being used. This can include sounds from traffic, construction, or other people talking. Background noise which is present in the recording in addition to the speaker's voice includes noises from the recording equipment or from the surrounding environment. Electrical interference can occur when the audio signal is contaminated by other electrical signals, such as from power lines or other electronic devices. The recording equipment

itself can also introduce noise into the audio signal. This can happen due to limitations in the equipment or due to improper setup or calibration.

Although deep neural networks improved the performance of speaker verification systems, noise interference still affects them, making robustness a more important metric and characteristic. Le Prell and Clavier, 2016 looked into how various background noises affected the speech signal. They suggested in their work that background noise may obscure (or mask) the details of another sound, such as the speaker's voice.

The additive noise, which is an elementary distortion model of a target signal with an unwanted signal, can be modelled as below;

$$s(t) = x(t) + n(t) \tag{2.9}$$

Babble noise is another type of additive noise that is characterized by multiple people speaking simultaneously, producing a chaotic mixture of voices. It is often described as a "babbling brook" of human voices. Babble noise is typically seen as one of the most challenging types of noise to deal with, as it can significantly degrade the quality of speech signals and make it difficult to distinguish individual voices. Babble noise can occur in a variety of settings, such as crowded public spaces, classrooms, or other noisy environments where many people are speaking at the same time.

Le Prell and Clavier, 2016 divided the noise types into two categories: stationary and non-stationary. The stable spectral and temporal properties of features, such as the noise from an idle car, are termed stationary noise. Stationary noise can be problematic in speech and audio processing, as some of the speaker's words may be obscured by this noise which can mask important speech information. The second sort of noise is non-stationary noise, like a dog barking. This type of noise has two different components: spectral fluctuation, in which the frequency components fluctuate rapidly over time, and temporal envelope variation, in which the noise level changes with time.

Zhao et al., 2014, and Ming et al., 2007 studied the impact of different noise environments on the DNN model's performance. Ming et al., 2007 experimented with various noise types, including engine noise, restaurant noise, and pop song noise, which are artificially mixed with the clean signal at different signal-to-noise ratios (SNRs) ranging from 10 dB to 20 dB.

Reverberation

Reverberation is the persistence of sound in an enclosed or semi-enclosed space after the sound has been produced. It results from sound waves reflecting off surfaces such as walls, ceilings, and floors and then bouncing back and forth until they gradually lose energy and fade away. Surfaces like walls, mirrors, and tables reflect, refract, and absorb these waves. The amount of reverberation in a space depends on the size and shape of the room, the materials and surfaces present, and the properties of the sound source. A figurative illustration of reverberation is shown in Figure 2.10.

A microphone records direct and attenuated (reflected) waves at various times. The room impulse response (RIR) is the transfer function between the sound source and the microphone. It includes the effects of reflections, diffraction, and absorption of sound waves within the room. The RIR is influenced by the room's characteristics, the source position, and the microphone position. The RIR is a finite impulse response filter used with a large number of taps (in the thousands) to achieve high accuracy in simulating complex room acoustics. The longer the FIR filter, the more reflections and details of the

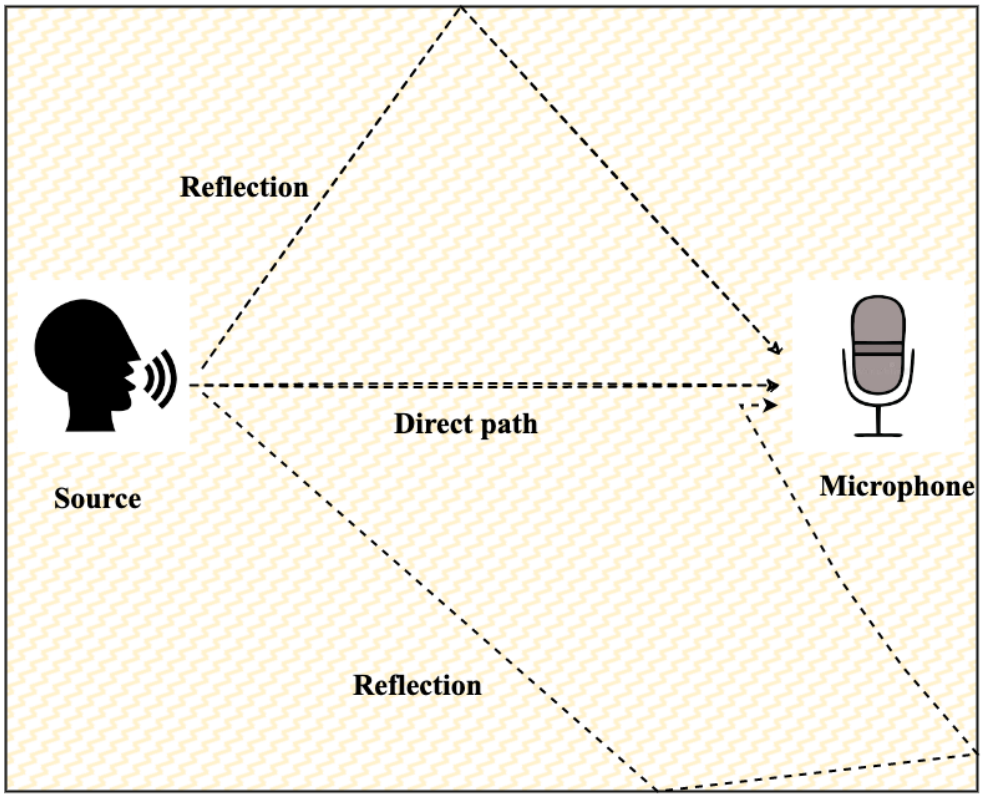

Figure 2.10: Graphical illustration of reverberation.

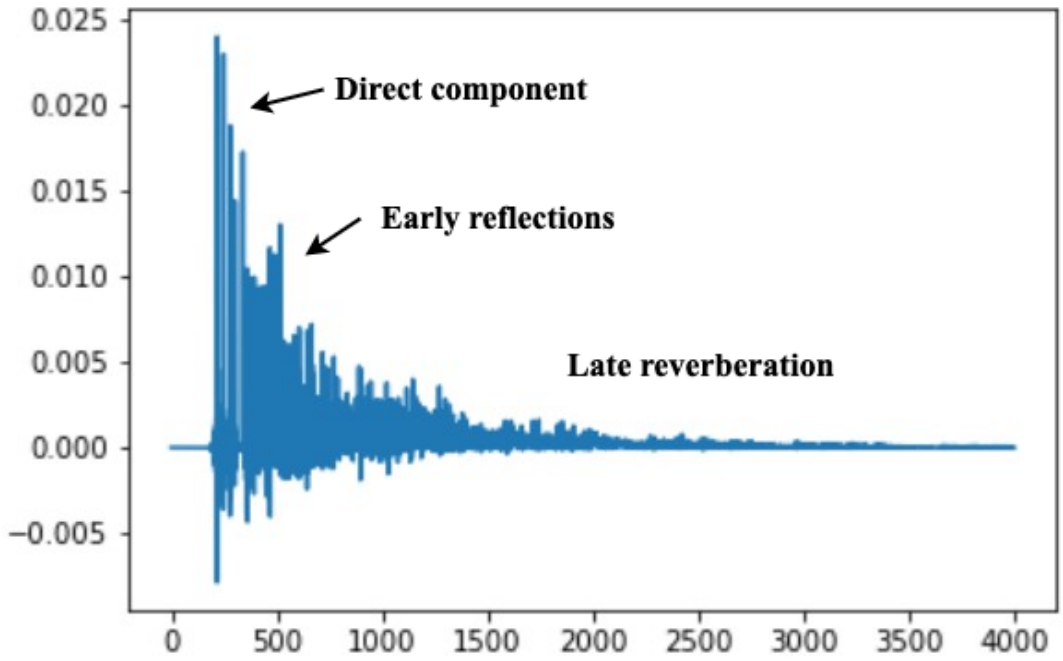

Figure 2.11: Plot of Room Impulse Response.

room acoustics are captured, providing a more precise representation of the RIR, which is essential for creating convincing spatial audio effects, such as the perception of distance, direction, and immersion. However, a longer FIR filter also requires more computational

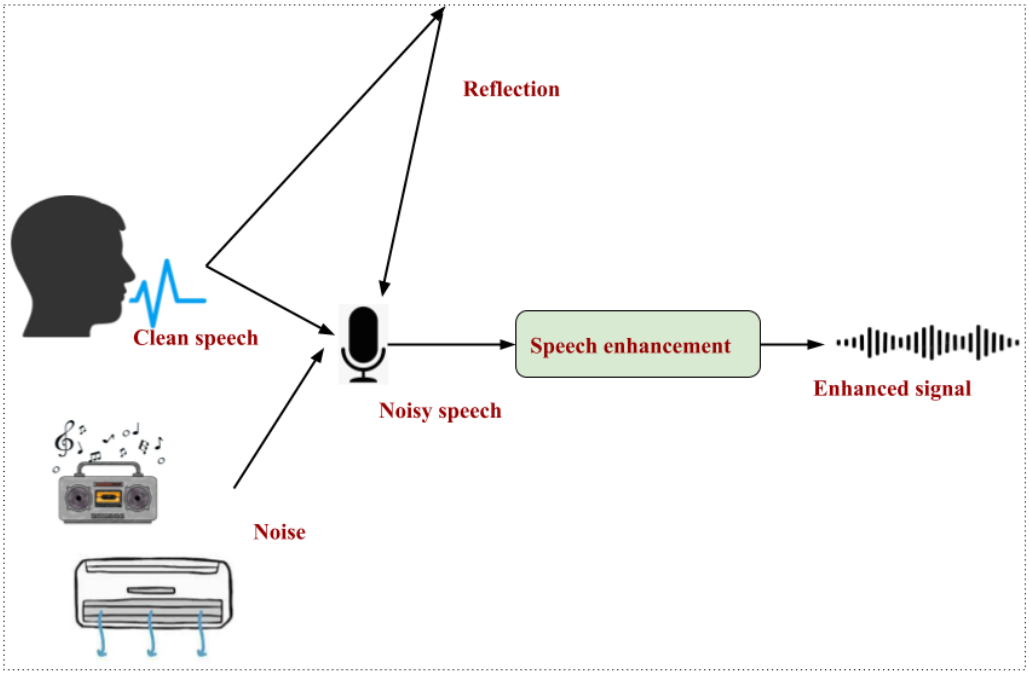

Figure 2.12: A graphical representation of generic speech enhancement process.

resources and memory to process. Therefore, the choice of the length of the FIR filter depends on a trade-off between the desired accuracy and the available computational resources.

By convolving an audio signal with an RIR represented as an FIR filter, we can simulate the effect of sound propagation within a room. In this context, a room is characterized by its reverberation time (RT60). The reverberation time of a room or space is defined as the time required for the sound to decay. It is often difficult to measure the RT60 accurately as it is almost impossible to generate a consistent and stable sound level.

2.5 Speech enhancement

As discussed in section 2.4.4, speech signals are corrupted by various factors, such as noise or reverberation, which degrades the speaker verification performance. Speech enhancement can be used as a pre-processing to improve speaker verification performance in noisy reverberant conditions. According to Bai and Zhang, 2020, speech enhancement is a natural choice for alleviating the noisy effects of the input speech signal as it aims to improve the quality and intelligibility of the speech signal by reducing or removing the unwanted noise and distortions that can affect its perception. A graphical illustration of the generic process of speech enhancement is shown in Figure A.2.

Some studies on speech enhancement focus on the problem of speech corruption caused by additive environmental noise commonly referred to as the noise reduction problem. Even though this subject has been thoroughly studied over the past few decades, noise reduction is still challenging for several reasons. Since the characteristics of the noise change over time, it is challenging to develop algorithms that can effectively remove it without damaging the desired signal. In some cases, it may be difficult to know beforehand the type and characteristics of the noise present in the input signal. This is

particularly challenging in scenarios where multiple noise sources are present or the noise is highly variable. In such cases, it is difficult to develop a noise reduction algorithm that can effectively remove the noise without distorting the desired signal. The signal-to-noise ratio (SNR) of the input signal can significantly impact the performance of noise reduction algorithms. When the SNR is low, the desired signal is heavily corrupted by noise, making it challenging to differentiate between the two. In such cases, the noise reduction algorithm may remove not only the noise but also parts of the desired signal, resulting in speech distortion and reduced intelligibility. This makes it considerably more challenging to create speech-processing algorithms that work well in various environments and conditions.

Speech enhancement can be mono- or multi-channel, depending on the number of microphones. We give in the next section a brief analysis of the single-channel speech enhancement in the following section. Since the main focus of this thesis is multi-channel speech enhancement, Section 2.5.2 gives a detailed review of the different approaches applied for multi-channel speech enhancement.

2.5.1 Single-channel speech enhancement

Single-channel speech enhancement is used to improve the quality and intelligibility of speech signals that are recorded with a single microphone in the presence of noise and reverberation. This is achieved by exploiting the statistical properties of the speech and noise signals, such as their spectral characteristics, temporal correlation, and sparsity. Single-channel speech enhancement techniques typically operate in a transformed domain, such as the frequency domain, where the signal and noise components can be separated and processed independently.

Recent approaches have considered speech enhancement as a supervised learning problem [Weninger et al., 2015, Park and Lee, 2016, Pascual et al., 2017]. The clean speech signal or other target features are thus estimated using a machine learning algorithm trained using a data-driven approach. The supervised speech enhancement approaches have benefited from DNN's rapid advancement, which has produced amazing achievements in various technological fields, including speech and signal processing. DNNs have demonstrated successful single-channel speech enhancement outperforming traditional statistical methods (like spectral subtraction, Wiener filtering, non-negative matrix factorization and Hidden Markov Models and establishing themselves as state-of-the-art algorithms. These techniques have moved towards using powerful network architectures that consider all areas of the processing pipeline, from important feature extraction to reconstructing the time signal. DNNs are capable of modelling non-linear relationships between the input and output signals and can automatically extract relevant features from the input signal without relying on hand-crafted features like MFCCs. DNN-based approaches focus on improving performance in increasingly difficult situations with low SNRs, non-stationary noises or interfering sources, and real-time applications.

However, single-channel speech enhancement methods are signal-adaptive frequency filters as they operate on the noisy speech signal to estimate a filter that is specific to the spectral properties of the signal. They adjust the filter coefficients based on the local characteristics of the noisy speech signal, such as the noise level and spectral properties. Single-channel speech enhancement approaches typically struggle to minimize background noise without introducing noticeable artifacts (such as musical noise [Cappe, 1994]) or speech distortion because the speech and noise signals typically occupy overlapping fre-

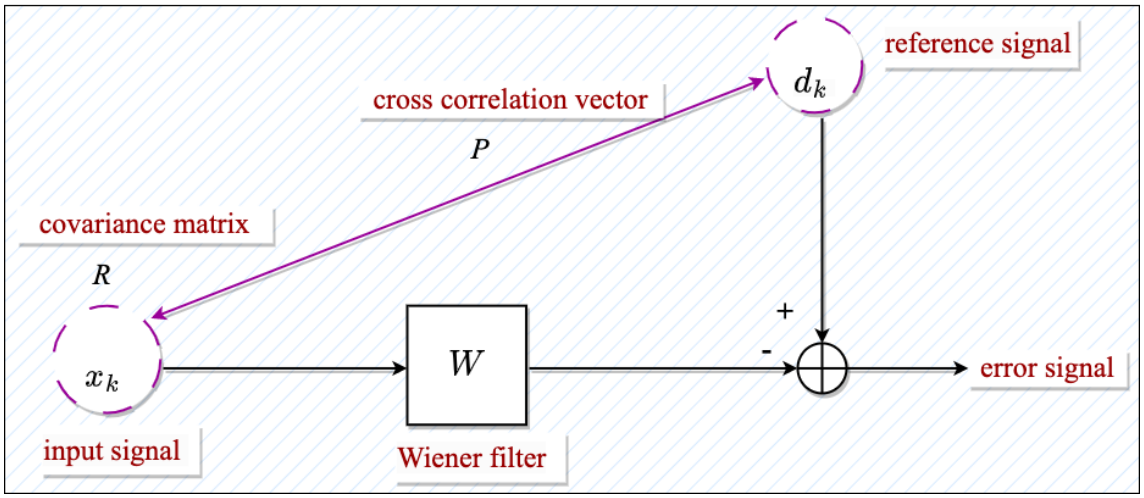

Figure 2.13: General illustration of conventional Wiener filter.

quency bands.

Spectral subtraction

The spectral subtraction technique [Boll, 1979] is probably one of the earliest methods used for speech enhancement. The basic idea behind this technique is to estimate the noise power spectral density (PSD) from the noisy speech signal and then subtract the estimated noise PSD from the noisy speech PSD to obtain an estimate of the clean speech PSD. The clean speech signal can then be obtained by inverse Fourier transform of the clean speech PSD.

The spectral subtraction technique consists of (i) Windowing: The noisy speech signal is divided into overlapping frames of a fixed length, and a window function is applied to each frame to reduce spectral leakage. (ii) Estimating the noise PSD: The PSD of the noise is estimated from the noisy speech signal using a method such as averaging the PSDs of silent frames or a voice activity detection (VAD) algorithm. (iii) Subtraction: The estimated noise PSD is subtracted from the PSD of the noisy speech signal to obtain an estimate of the PSD of the clean speech signal. The amount of subtraction can be controlled using a scaling factor, which is typically determined empirically. (iv) Reconstruction: The estimated clean speech PSD is inverse Fourier transformed to obtain the corresponding clean speech signal.

While the Spectral Subtraction technique can be effective at reducing noise in speech signals, it has some limitations. One limitation is that it assumes the noise is stationary and does not change over time. This can be problematic if the noise is non-stationary, as the estimated noise PSD may not accurately represent the true noise PSD. Additionally, the technique can introduce distortion and artefacts in the reconstructed speech signal.

Wiener filter

The Wiener filter is a widely used digital signal processing technique for noise reduction and signal enhancement. Wiener filter [Wiener et al., 1949] is a statistically optimal criterion with well-defined statistical assumptions that has become one of the broadly used solutions to system identification issues and has significant applicability across a

wide domain [Benesty and Huang, 2013, Ljung, 1998, Haykin, 2002].

The Wiener filter is a linear filter that is used to estimate a desired signal from a noisy or degraded signal. The filter works by minimizing the mean squared error between the estimated signal and the desired signal. The Wiener filter is designed based on the statistical properties of the desired signal and the noise in the input signal. The basic idea behind the Wiener filter is to estimate the power spectral density (PSD) of the desired signal and the noise in the input signal. The filter then applies a frequency-dependent gain to the input signal, which is determined based on the ratio of the PSD of the desired signal to the total PSD of the input signal (i.e., the sum of the PSDs of the desired signal and the noise).

The Wiener filter can be implemented in both the time domain and the frequency domain. In the time domain, the filter is typically implemented using a finite impulse response (FIR) filter. In the frequency domain, the filter is implemented using the Fourier transform of the input signal and the filter coefficients, which are calculated based on the estimated PSDs.

The gain function of the Wiener filter in the STFT domain can be computed by solving the following minimization problem;

$$G_{WF}(t,f) = \underset{G(t,f)}{argmin}\mathbb{E}\{|X(t,f) - G(t,f)Y(t,f)|^2\} \tag{2.10}$$

Thus, the Wiener filter minimizes the mean square error between the clean and estimated speech spectra. Considering that the noise and clean speech STFT coefficients are uncorrelated zero-mean random variables, and their second-order variance is ${\sigma_x}^2(t,f)$ and ${\sigma_n}^2(t,f)$, Wiener filter gain can be obtained as;

$$G_{WF}(t,f) = \frac{\xi(t,f)}{\xi(t,f)+1} \tag{2.11}$$

where $\xi(t,f) = {\sigma_x}^2(t,f){\sigma_n}^2(t,f)$ is the priori SNR. As can be seen, the gain function is close to zero for bins with a low SNR, i.e., noise-dominant bins, suppressing the frequencies, while the gain function is close to one for bins with high SNR, i.e., speech-dominant bins.

Apart from the conventional Wiener filter, there are other Wiener-like functions have been proposed, such as square-root or parametric Wiener filter [Loizou, 2007], the codebook-driven filter [Srinivasan et al., 2006], and the application of the psychoacoustic characteristics of the human auditory system [Hu and Loizou, 2003, Hu and Loizou, 2004].

Ideal time-frequency masks

Time-frequency mask estimation is a commonly used technique in speech enhancement to separate speech from background noise. The goal of the technique is to estimate a time-frequency mask that can be applied to the noisy speech signal to attenuate the background noise while preserving the speech signal.

The time-frequency mask is typically computed using a two-step process. The first step involves computing a spectrogram of the noisy speech signal using a short-time Fourier transform (STFT). The second step involves estimating a time-frequency mask that can be applied to the noisy signal to separate speech from background noise. One common approach is to use a binary mask, where values of 1 indicate that the corresponding time-frequency bin should be preserved (i.e., the speech signal is present), and values

Masking approach	Expressions
IBM	$M_x(t,f) = \{^{1}_{0} \frac{\lvert X(t,f)\rvert^2}{\lvert N(t,f)\rvert^2+\lvert X(t,f)\rvert^2}$
IRM	$M_x(t,f) = (\frac{\lvert X(t,f)\rvert^2}{\lvert X(t,f)^2\rvert+\lvert N(t,f)^2}$
cIRM	$M_x{}^r(t,f) = \frac{Y^r(t,f)X^r(t,f)+Y^i(t,f)X^i(t,f)}{Y^{2,r}(t,f)+Y^{2,r}(t,f)}$
	$M_x{}^i(t,f) = \frac{Y^r(t,f)X^i(t,f)-Y^i(t,f)X^r(t,f)}{Y^{2,r}(t,f)+Y^{2,i}(t,f)}$

Table 2.1: Figurative representation of the three ideal ratio masks. The subscripts r and i represent the real and imaginary parts, respectively.

of 0 indicate that the corresponding time-frequency bin should be attenuated (i.e., the background noise is present).

Ideal ratio masks (IRM) and complex ratio masks (cIRM) are other types of time-frequency masks. These masks are used in evaluations to get some comparison corresponding to using a perfect mask or as target values in training processes. They differ in their approach to estimating the mask and the information they use to compute it. The IRM is a real-valued mask representing the ratio of the clean speech power to the noisy speech power in each time-frequency bin. Unlike the ideal binary mask (IBM), the IRM does not assume that the noise is additive or stationary and can adapt to changes in the noise characteristics. However, the IRM assumes that the speech and noise are statistically independent, which may not always be true.

The cIRM is a complex-valued mask that represents the ratio of the complex Fourier coefficients of the clean speech signal to the complex Fourier coefficients of the noisy speech signal in each time-frequency bin. The cIRM can capture phase information, which is important for preserving the quality of the speech signal. However, computing the cIRM requires estimating the phase of the clean speech signal, which can be challenging in practice. Table 2.1 presents the mathematical expression for IBM, IRM and cIRM. The choice of time-frequency mask depends on the specific characteristics of the speech and noise signals and on the desired trade-off between noise reduction and speech distortion.

The time-frequency mask can be estimated using various methods, including Wiener filtering, subspace methods, and deep learning-based approaches. For example, a deep neural network can be trained to predict the time-frequency mask from the noisy and desired clean speech signals. The estimated mask can then be applied to the noisy speech signal to obtain the enhanced speech signal.

2.5.2 Multi-channel speech enhancement

Multi-channel speech enhancement is a signal processing technique to obtain a clean signal from noisy mixtures recorded with multiple microphones. The development of multi-channel speech enhancement techniques was made possible by the advent of devices having arrays of two or more microphones. When the speech signal and the background noise are located in different positions in the room, the desired signal can be spatially separated from the noise since the speech signal, and background noise is located in different positions in the room [Van Trees, 1969]. The multi-channel approach enables us to take advantage of the spatial information in the signals, such as the location of the target speaker, the acoustic channel's characteristics, or the statistical characteristics

of the noise field across the microphones. Using multiple microphones, the multi-channel approach can capture the sound from different spatial positions and estimate the direction of arrival of the desired speech signal and the interfering noise sources. If the target speaker and the interfering sources are spatially separated, then the multi-channel approach can effectively separate the two signals and attenuate the noise while preserving the quality of the speech signal. However, if the target speaker and the interfering sources are closely located, such as in a reverberant room, then the multi-channel approach may not be as effective in separating the two signals and may result in some alteration of the speech signal.

Today's smart devices using speech-based applications are equipped with multiple microphones for exploiting spatial features and increasing speech intelligibility. With the availability of multiple microphones, speech-based applications hope to provide good intelligibility and advance the technology of mobile devices even in a noisy-reverberant environment. In the following sections, we explain some multi-channel speech enhancement techniques.

Beamforming

Beamforming is often used in conjunction with a microphone array and other signal processing techniques, such as spatial and temporal filtering, to perform spatial filtering. To distinguish the desired components from the background noise, reverberation, and interfering signals, beamforming relies on spatial and spectral information. The beamformer aims to extract the signal originating from the region of interest and attenuate signals from different locations.

The noisy speech signal at each microphone can be represented as;

$$x_k(t, f) = s_k(t, f) + n_k(t, f), \tag{2.12}$$

where $k = 1, ..., K$ is the microphone index and k represents the number of microphones. Let us also assume that the clean speech signal at each microphone is different due to the room acoustics and the microphone responses.

There are two main categories of beamformers: data-independent and data-dependent, which are described in the following sections.

Data-independent beamformers

The filter coefficients are pre-determined to extract the desired signal regardless of the statistical characteristics of the source signals in a data-independent beamformer design. The objective is to achieve spatial focus on the desired source that originated from the region of interest and to suppress undesirable signals like interference or background noise. The key benefits of Data-independent beamformers include the ability to prevent signal distortion without the need for controlling algorithms, and their comparatively low numerical complexity and simplicity of implementation [Chen et al., 2007]. However, because they are not well adaptable to changing acoustic surroundings, fixed beamformers have a limited ability to suppress noise.

Delay-and-sum [Veen and Buckley, 1988] is one of the examples of data-independent beamformers, which we briefly explain in the following section.

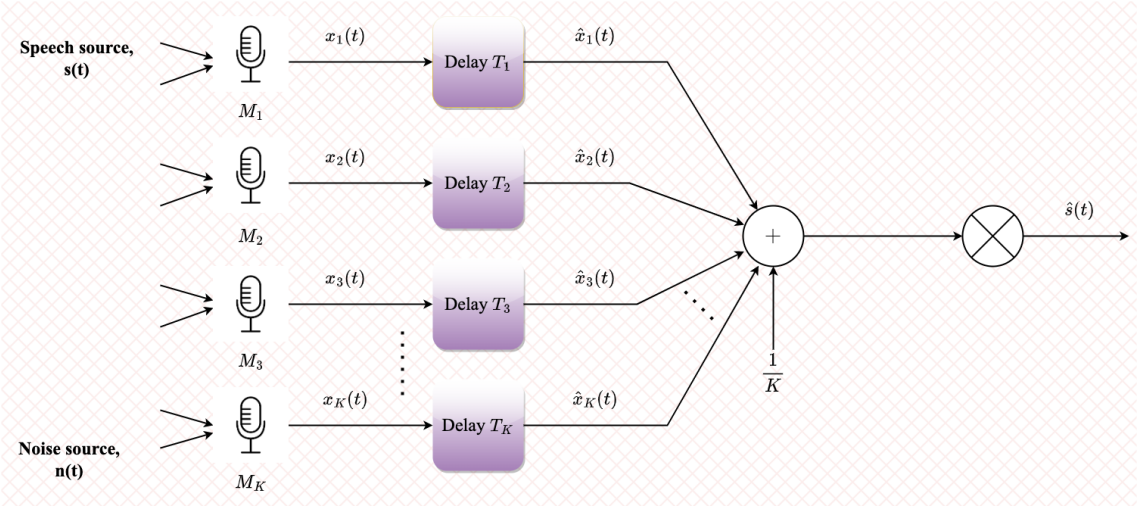

Figure 2.14: Graphical illustration of the structure of delay-and-sum beamformer.

Delay-and-Sum Beamformer

The delay-and-sum beamformer combines the signals from multiple microphones to maximize the desired signal while minimizing noise and interference from other directions. The technique works by applying a set of delays and weights to the signals from each microphone before summing them together. As they add up together coherently, the desired signal components are reinforced, and the noise is eliminated as they are added together destructively.

Figure 2.14 illustrates the structure of the delay-and-sum beamformer, where the signal arriving at the microphone is the combination of both the desired signal and the noise signal. Delay-and-sum can be expressed as follows;

$$x_l(t) = s_l(t) + n_l(t) \tag{2.13}$$

The delayed signal can be expressed as follows;

$$\hat{x}_l(t) = s_l(t - T_l) + n_l(t - T_l) \tag{2.14}$$

where T_l is the delay to the l_{th} signal. The aligned signals are added together to form the beamed output $\hat{s}(kT)$. The output signal is expressed as follows;

$$\hat{s}(kT) = \frac{1}{N}\Sigma_{l=1}^{N}\hat{x}_l(kT) + \frac{1}{N}\Sigma_{l=1}^{N}\hat{n}_l(kT) \tag{2.15}$$

Data-dependent beamformers

The statistical characteristics of the incoming signals constitute the basis for the design of a data-dependent beamformer, also known as an adaptive beamformer. In contrast to data-independent beamformers, an adaptive beamformer uses spatial filtering. Compared to data-independent beamformers, this usually results in greater noise suppression performance. Additionally, such a design is regularly updated to track current statistics regarding the statistics of the room propagation conditions. The linearly constrained minimum variance (LCMV) beamformer [Frost, 1972] is one of the well-known examples of an adaptive beamformer. The LCMV beamformers' core principle is to apply a linear

constraint on the weight vector to control the beamformer output and keep the intended signals' gain and phase constant. This aids in maintaining the target signal and reducing the contribution of interference and noise signals [Van Veen and Buckley, 1988]. The generalized sidelobes canceller (GSC) [Griffiths and Jim, 1982], a different unconstrained design problem of the LCMV beamformer, is another example of an adaptive beamformer.

Since adaptive beamforming techniques can adjust to shifting acoustic environments, they typically outperform data-independent beamforming techniques at removing noise. They exhibit speech distortion and cancellation as a result of modelling flaws and due to their extreme sensitivity.

In the following sections, we describe some of the data-dependent beamformers we have used in this thesis work.

Minimum variance distortionless response beamformer

The Minimum Variance Distortionless Response (MVDR) beamformer [Capon, 1969] is a data-dependent beamforming technique designed to reduce noise at the output of the beamformer while imposing a distortionless constraint on the target speech signal. MVDR is also known as Capon filter [Capon, 1969] and is perhaps the most popularly used super-directive beamformer. The main goal of the MVDR beamformer is to minimize the variance of the recorded signal. The MVDR filter can be obtained by solving the following optimization problem;

$$\mathbf{w}_f^{MVDR} = \underset{\mathbf{w}f}{argmin} \boldsymbol{E}|w_f^H \mathbf{u}_{t,f}|^2, \tag{2.16}$$

subject to

$$\mathbf{w}_f^H \mathbf{h}_f = 1, \tag{2.17}$$

where $\{\boldsymbol{E}|w_f^H \mathbf{u}_{t,f}|^2 = \mathbf{w}_f^H \mathbf{R}_{u,f} \mathbf{w}_f\}$ is the power spectrum density of the output noise signal and $\mathbf{R}_{\mathrm{u,f}} = \boldsymbol{E}\mathbf{u}_{\mathrm{t,f}}\mathbf{u}_{\mathrm{t,f}}^{\mathrm{H}}$ is the spatial correlation matrix of the noise signals. If we assume that the target speech and noise signals are uncorrelated, we can express the power spectrum density of the output noise as;

$$\begin{aligned} \mathbf{E}\ |\mathbf{w}_f^H \mathbf{u}_{t,f}|^2 &= \mathbf{w}_f^H \mathbf{R}_{y,f} \mathbf{w}_f - \mathbf{w}_f^H \mathbf{R}_{o,f} \mathbf{w}_f & (2.18)\\ &= \mathbf{w}_f^H \mathbf{R}_{y,f} \mathbf{w}_f - \mathbf{w}_f^H h_f \phi_X \mathbf{w}_f h_f^H & (2.19)\\ &= \mathbf{w}_f \mathbf{R}_{y,f} \mathbf{w}_f - \phi_X & (2.20) \end{aligned}$$

where ϕ_X is the power spectral density of the target speech and $\mathbf{R}_{y,f} = E\{|y_{t,f}|^2\}$ and $\mathbf{R}_{o,f} = E\{|\mathbf{o}_{t,f}|^2\}$ are the spatial correlation matrices of the microphone signals and the source. The distortionless constraint means that the second term does not depend on $\mathbf{w}_f$. Consequently, the optimization problem of equation (2.14) can be reformulated as follows;

$$\mathbf{w}_f^{MVDR} = \underset{\mathbf{w}f}{argmin} \boldsymbol{E} w_f^H \mathbf{R}_{y,f} \mathbf{w}_f, \tag{2.21}$$

subject to

$$\mathbf{w}_f^H \mathbf{h}_f = 1, \tag{2.22}$$

By solving this optimization problem, we get the following expression for the MVDR filters;

$$\mathbf{w}_f^{MVDR} = \frac{\mathbf{R}_{t,f}^{-1}\mathbf{h}_f}{\mathbf{h}_f^H \mathbf{R}_{t,f}^1 \mathbf{h}_f}. \tag{2.23}$$

To be noted, to compute the filters, we first need to estimate the steering vector $\mathbf{h}_f$ and the spatial correlation matrix of the microphone signals.

MVDR involves the estimation of the noise covariance matrix. As MVDR is an adaptive beamformer, it can adapt to a noisy environment for maximum noise reduction. This is accomplished by re-calculating the filter coefficients using the cross-spectral density matrix of noise.

Generalized eigenvalue beamformers

The Generalized eigenvalue (GEV) beamformer operates by finding the optimal set of weights to apply to each of the sensors in the array in order to maximize the output signal-to-noise ratio.

The GEV beamformer is based on the idea that the signal received by an array of sensors can be modelled as a linear combination of the desired signal and the noise. By finding the optimal set of weights for each sensor in the array, the GEV beamformer can effectively filter out unwanted noise and interference while amplifying the desired signal.

The GEV beamformer first constructs a covariance matrix from the data received by the array of sensors. This covariance matrix represents the statistical correlation between each pair of sensors in the array. The GEV beamformer then performs a generalized eigenvalue decomposition on the covariance matrix, which produces a set of eigenvectors and eigenvalues. The eigenvectors represent the optimal set of weights to apply to each of the sensors, while the eigenvalues represent the strength of the corresponding signal. Once the eigenvectors have been determined, the GEV beamformer applies these weights to the signals received by each sensor in the array. The resulting output signal has a higher signal-to-noise ratio than the original input signal, which makes it easier to identify and analyze the desired signal.

Along with the GEV beamformer, a Blind Analytic Normalization (BAN) postfilter [Warsitz and Haeb-Umbach, 2007] is used. The GEV-BAN obtains better performance than the MVDR beamformer. Beamforming combined with post-filtering is a popular method for multi-channel speech enhancement.

Multi-channel Wiener filter

The multi-channel Wiener filter (MWF) can be considered a type of beamformer since it realizes the multi-channel filtering of microphone signals to reduce noise. The MWF is the solution to a Minimum Mean Square Error (MMSE) problem that allows for noise reduction. MWF estimates the speech component of a reference signal at the microphone(s).

The MWF incorporates the spatial properties of the signals into the filter design. The spatial properties of the signals can be characterized by the inter-channel correlation and the direction of arrival of the target signal. These spatial properties are used to estimate the spatial covariance matrix, which is then incorporated into the filter design to improve the enhancement performance.

The MWF can be generated as a multi-channel MMSE linear filter like its single-channel counterpart. The filter that minimizes this cost function is the MWF [Widrow et al., 1967, Doclo and Moonen, 2002a], which is the optimal solution in the mean squared

sense, i.e. it minimizes the MSE between the desired signal s_i and the estimated signal. The MWF solution is defined by;

$$\mathbf{w}_{MWF} = \underset{\mathbf{w}}{argmin}\mathbb{E}\{|s_1 - \mathbf{w}^H\mathbf{y}|^2\} \tag{2.24}$$

where $\mathbb{E}$ is expectation operator and H is Hermitian transpose.

Solving equation 2.24 yields;

$$\mathbf{w}_{MWF} = {R_y}^{-1}R_{ys}e_1 \tag{2.25}$$

where Ry is the correlation matrix of the input signal, R_{ys} is the cross-correlation matrix between the input signal and its speech component, and with $e_1 = [1, 0, ...0]$ the vector which selects the (first) reference channel. If the noise and the speech are uncorrelated and of zero mean, we have $R_{ys} = R_s$ and $R_y = R_S + R_n$. The equation 2.25 is equivalent to;

$$\mathbf{w}_{MWF} = (R_s + R_n)^{-1}R_se_1 \tag{2.26}$$

By factoring equation 2.24, it can be shown that the MWF is equivalent to an MVDR followed by a single-channel filter if the noise follows the Gaussian distribution [Simmer et al., 2001, Balan and Rosca, 2002].

The MWF has the advantage of not relying on the knowledge of the angles of arrival. A priori knowledge of the position of microphones and sources is not required. It requires the knowledge of noise only period, which can be obtained with VAD. Equation 2.24 can be declined in different variants, allowing for adjustment of certain criteria or making a more robust filter implementation.

Rank-1 multichannel Wiener filter

When only a speech source is present, in the absence of reverberation, it can be assumed that the speech signal, as measured by the microphones, depends on vector only:

$$s = s_{src}d, \tag{2.27}$$

where s_{src} is the non-reverberant speech signal. Thus the spatial correlation matrix of speech can be calculated as;

$$\begin{aligned} R_s &= \mathbb{E}\{ss^H\} && (2.28)\\ &= \mathbb{E}\{s_{src}dd^Hs^*_{src}\} && (2.29)\\ &= {\sigma_s}^2dd^H. && (2.30) \end{aligned}$$

where $\sigma_s^2 = \mathbb{E}\{s_{src}s^*_{src}$ is the speech power spectral density.

By re-enacting equations 2.26 and (2.27) using the Woodbury identity [Woodbury, 1950], we finds the following rank 1 Wiener filter [Doclo et al., 2006, Souden et al., 2010] expression;

$$\mathbf{w} = \frac{{R_n}^{-1}R_se_1}{1 + tr({R_n}^{-1}R_s)} \tag{2.31}$$

where tr denotes the trace operator. Note that since $R_n{}^-1R_s$ is of rank-1, the trace of this matrix is equal to its only eigenvalue.

The advantage of this formulation, strictly equivalent to the formulation in 2.26 The matrix R_s is of rank 1. In practice, we force the rank of R_s by taking $R_s = aa^H$ with the eigenvector of R_s corresponding to its eigenvalue. We are thus assured that R_s is well determined [Cornelis et al., 2011].

Speech distortion-weighted multi-channel Wiener filter

If noise and speech are uncorrelated, the cost function in 2.24 can be decomposed into two terms;

$$J_{(\mathbf{w})} = \mathbb{E}\{|s_1 - \mathbf{w}^H\mathbf{s}|^2\} + \mu\mathbb{E}\{\mathbf{w}^H\mathbf{n}|^2\}. \tag{2.32}$$

The parameter $\mu \in R^+$ makes it possible to weight the noise reduction (expressed in the term $\mathbb{E}|\mathbf{w}^H n|^2$) by speech distortion (expressed in the term $\mathbb{E}\{|s_1 - \mathbf{w}^H s|^2\}$). Choosing a high μ amounts to gives more importance to reducing noise, even if it distorts the speech signal. On the other hand, taking a weak μ ensures minimal speech distortion, even if it means filtering less noise. Thus, in the extreme case $\mu = 0$, for which no distortion is applied to the speech signal, it gives the same solution as the MVDR.

The vector $\mathbf{w}$ that minimizes 2.32 is the distortion-weighted multichannel Wiener filter (speech distortion weighted multichannel Wiener filter (SDW-MWF)) [Doclo and Moonen, 2002a], can be expressed as follows;

$$\mathbf{w}_{SDW-MWF} = (\mathbf{R}_s + \mu\mathbf{R}_n)^{-1}\mathbf{R}_s e_1 \tag{2.33}$$

The MWF and the SDW-MWF are equivalent for $\mu = 1$. Moreover, [Doclo et al., 2010] demonstrate the SDW-MWF can be decomposed into an MVDR by applying a gain. In the same way, as for the MWF, we can assume rank 1 speech correlation matrix, which gives the following rank 1 SDW-MWF expression;

$$\mathbf{w}_{r_1 SDW-MWF} = \frac{\mathbf{R}_n{}^{-1} R_s e_1}{\mu + tr\{R_n{}^{-1} R_s\}} \tag{2.34}$$

2.6 DNN-based speech enhancement

The speech enhancement approaches discussed above are mainly based on statistical signal processing. These methods were developed based on assumptions about the correlations, stationarity, and other statistics of the clean speech signal and the environmental noise. Based on these assumptions, a mathematical formulation for the estimate of the clean speech signal can be developed, producing closed-form estimators that are either ideal or approximations that are suboptimal and provide good enough results. Although implementing these techniques into practical applications, their effectiveness depends heavily on the accuracy of the assumptions and the estimation of the underlying parameters. When the assumed characteristics of the speech and noise signal are maintained, traditional speech enhancement approaches, such as spectral subtraction [Boll, 1979], Wiener filter [Wiener et al., 1949], short-time spectral amplitude estimators [Ephraim and Malah, 1984], and maximum-likelihood spectral amplitude [McAulay and Malpass, 1980] algorithms have demonstrated good noise suppression performance. However, traditional approaches often introduce distortions into the enhanced speech signal in low SNR conditions, the presence of non-stationary noises, reverberant acoustic environments, etc. This

is mainly because effectively estimating these conditions' speech and noise properties is difficult.

Recently, DNN-based approaches have been widely used for speech enhancement. DNN-based approaches can more effectively characterize nonlinear mapping between noisy and clean signals than conventional approaches. This is mainly because they often comprise several layers of non-linear transformations between the observable data and the target to be estimated. In multi-channel scenarios, DNNs are applied to compute the time-frequency masks separating speech and noise from a mixture signal and are then used to estimate the speech and noise covariance matrices for beamforming [Heymann et al., 2016]. The state-of-the-art speech enhancement systems use DNN-based algorithms, and a large amount of training data is recorded under various noisy and reverberated conditions to achieve noise robustness. Dufera and Shimamura, 2009 proposed a conventional method for removing a room impulse, in which the RIR filter is estimated using DNN. Speech separation has also been accomplished using DNNs [Tu et al., 2014]. A DNN-based autoencoder for speech enhancement was proposed by Xu et al., 2014a with optimization [Xu et al., 2014b].

This section gives a brief overview of DNN architectures for speech enhancement. We focus mainly on commonly used (i) Recurrent neural networks, (ii) Convolutional neural networks, and (iii) Generative models.

2.6.1 Recurrent neural networks

Recurrent neural networks (RNNs) are among the best choices for time sequence data as they can learn from sequential input data. Recurrent nets do not require explicit contextual information feeding because past inputs are used to make predictions. Hidden layers in an RNN compute activations utilizing inputs from a lower layer and its output value at a prior step in the sequence. Hence, the network outputs depend on the current data and data from the previous time instants. Similar to how a standard RNN uses past context, a reverse directional RNN uses future context. The two are combined to create a bidirectional RNN. Long-short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997] and gated recurrent unit (GRU) [Cho et al., 2014] are popular choices of units to design a recurrent layer.

Xiao et al., 2017 used RNN to estimate the time-frequency masks targeted toward reliable speech recognition applications. The approach involves directly minimizing the cost function for acoustic speech recognition while also training the RNN to estimate the speech and noise masks independently. Self-attention with a dual-path RNN for time-domain speech enhancement was also studied [Pandey and Wang, 2020c].

2.6.2 Convolutional neural networks

Convolutional neural networks (CNN) are popularly used in image and video processing, but they can also be applied to various speech processing tasks [Xu et al., 2015]. They have been evaluated for time-domain speech signals for enhancement [Fu et al., 2017]. The authors have employed convolutional layers to model the raw waveform. More specifically, they have employed a convolutional network model to enable each output sample to depend locally on the neighbouring input regions. A connection between these convolutional layers and finite impulse response filters in this context can be established using the convolutional layer to take advantage of short temporal correlations.

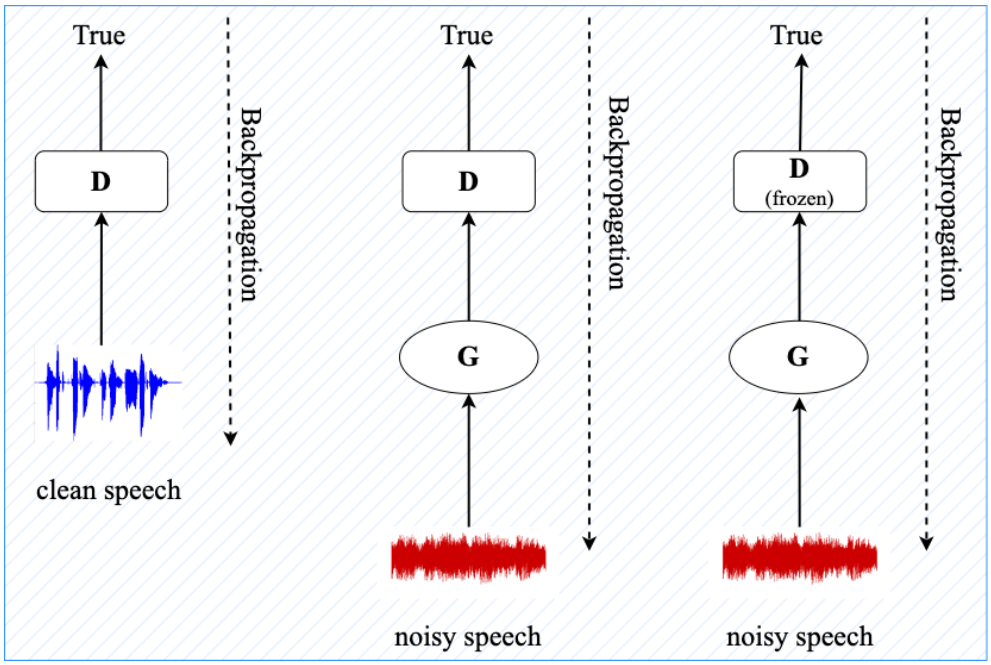

Figure 2.15: Illustration of adversarial training for speech enhancement. D refers to the discriminator, and G refers to the generator.

The convolutional layer can also operate in the time-frequency domain [Park and Lee, 2016] exploiting the temporal and frequency correlations of neighbouring bins. CNN was implemented for time-domain speech enhancement by Fu et al., 2017 where they proposed an alternative convolutional network architecture, namely a Redundant Convolutional Encoder-Decoder network to extract redundant representations of a noisy spectrum at the encoder and map it back to a clean spectrum at the decoder. Speech enhancement is further improved by using dilated convolutions [Rethage et al., 2017, Qian et al., 2017, Pandey and Wang, 2019]. Wavenet [Rethage et al., 2017] can predict target fields (a sequence of samples that represent a particular feature of the input signal, such as the fundamental frequency or the spectral envelope) instead of single samples, which reduces the time complexity and significantly improves the performance of the model. The convolutional nature of the model makes Wavenet flexible in the time dimension, resulting in supporting denoising variable-length audio. Dense connections [Pandey and Wang, 2020b], self-attention [Giri et al., 2019, Pandey and Wang, 2020a], and dual-path RNN [Luo et al., 2020, Pandey and Wang, 2020c] are some of the additional networks applied to improve the speech enhancement performance using CNN.

2.6.3 Generative models

Generative models have achieved promising performance in almost all data modalities by generating outputs that best resemble the source data. They can automatically learn the inherent natural features of a dataset irrespective of whether they are categories or dimensions, or something else. Generative adversarial network (GAN) [Goodfellow et al., 2014] simultaneously train two DNN models: a discriminator that calculates the likelihood that a sample came from the training data rather from the generator and a generator that captures the distribution of the training data. Figure 2.15 illustrates the adversarial training of GAN.

Generative models have also been used widely for speech enhancement by following a different paradigm where they learn the distribution of clean speech a priori. As a result, they aim to learn speech's fundamental characteristics, such as its spectral and temporal structure. With noisy or reverberant input signals considered outside the learned distribution, this prior knowledge can be used to generate clean speech. This concept is the foundation of several methods that use deep generative models to improve speech enhancement performance [Pascual et al., 2017, Bando et al., 2017, Leglaive et al., 2018, Richter et al., 2020, Bando et al., 2020].

Speech enhancement GAN (SEGAN) [Pascual et al., 2017] is an end-to-end speech enhancement approach. Speech signals are operated on the raw waveform rather than the spectral or higher-level domains. The noisy speech signal and the latent representation are the input in the SEGAN structure, and the generator's predicted output is the clean speech. Thus, the speech enhancement is carried out via the generator. Bando et al., 2017 proposed a probabilistic generative model called VAE-NMF, where they unified the probabilistic generative model of noisy speech spectra by combining a variational autoencoder-based generative model of speech spectra with a non-negative matrix factorization-based generative model of noise spectra for speech enhancement. Richter et al., 2020 proposed a generative approach to speech enhancement using a stochastic temporal convolutional network as a speaker-independent speech model to estimate the variance of clean speech, and estimation of the noise variance is based on a non-negative matrix factorization. A neural speech enhancement method with a statistical feedback mechanism based on a denoising variational autoencoder has been proposed for speech enhancement [Bando et al., 2020]. They train a denoising variational autoencoder consisting of two networks: the denoising encoder to estimate the latent vectors of clean speech from a noisy mixture and the generative decoder to generate a speech spectrogram from the latent variable. In the test time, they used the encoder output as a prior distribution of clean speech and updated the speech latent vectors to fit the input mixture signal, while a non-negative matrix factorization-based noise model estimated the noise signal.

Diffusion probabilistic models

Diffusion probabilistic models (DPM) [Ho et al., 2020, Sohl-Dickstein et al., 2015], a class of generative models are inspired by the non-equilibrium thermodynamics diffusion process and learn data-to-noise mapping in discrete steps. In thermodynamics, "diffusion" refers to the flow of molecular particles from high-density to low-density areas. DPMs learn to reverse the diffusion process to create desired data samples from the noise after defining a Markov chain of diffusion steps to gradually introduce random noise to data. Figure 2.16 shows the denoising diffusion process. Diffusion models are learned with a fixed procedure, and the latent variable has the same dimension as the original data. In the context of speech processing, DPMs were the first to introduce the modelling of complex data distributions using stochastic calculus [Chen et al., 2020].

Diffusion models work by corrupting the training data by adding Gaussian noise iteratively, then learning to recover the data by undoing this noise-adding process. After training, diffusion models can generate data by applying the learned denoising technique to randomly sampled noise. Diffusion probabilistic models (DPM) have shown impressive performance in image generation [Ho et al., 2020] and Text-to-speech systems [Jeong et al., 2021, Popov et al., 2021]. Recently, DiffuSE [Lu et al., 2021] and CDiffuSE [Lu et al., 2022] were proposed to recover clean speech signals from noisy signals based on Markov chains to provide a framework for diffusion probabilistic models. The Markov

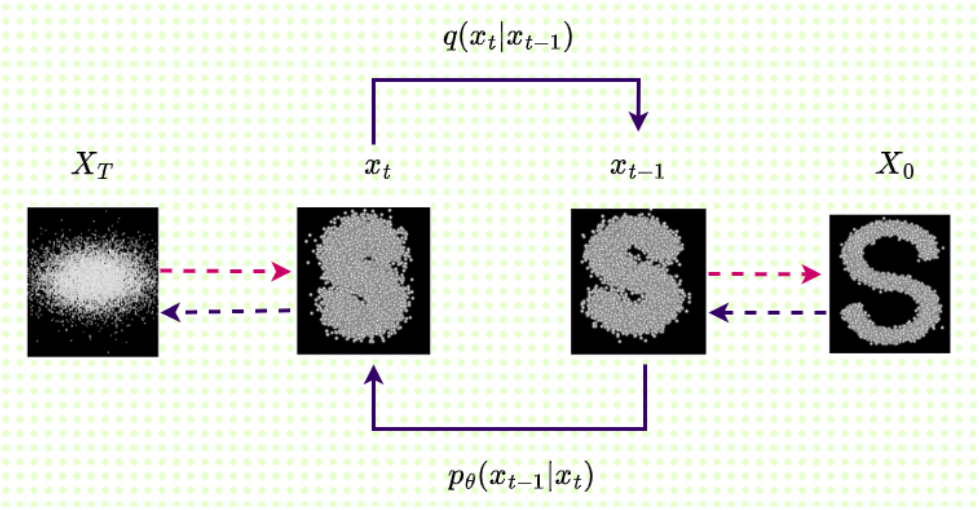

Figure 2.16: Graphical illustration of the process of Denoising Diffusion Probabilistic Model. The red arrow depicts the forward diffusion process, and the dark purple arrow depicts the reverse diffusion process.

chains are fixed during the model architectural design, resulting in the slower sampling of latent variables, thus leading to slower inference speed. The reason why Markov chains can lead to slower sampling is that they impose a fixed dependence structure between the latent variables. This means the transitions between states are predetermined and do not change during the model training or inference process compared to scoring-based DPM. As a result, the model may require more iterations to converge to the posterior distribution of the latent variables, which can slow down the inference speed. DiffuSE and CDiffuSE use fixed Markov chains for training diffusion models under the framework of DPM. In contrast, scoring-based diffusion probabilistic models do not rely on fixed Markov chains to model the dependencies between the latent variables. Instead, they use a score-based approach to update the posterior distribution of the latent variables, which can lead to faster convergence and faster inference speed.

Score-based generative models are a new direction in generative modelling that relies on learning score functions, which are gradients of log probability density functions. The models generate samples using Langevin-type sampling and have several advantages over other model families, such as GAN-level sample quality, flexible architectures, exact log-likelihood computation, and inverse problem-solving without re-training [Song et al., 2020b]. In a score-based DPM, the model learns to estimate the score function, which is the gradient of the log-likelihood with respect to the input data. This score function can be used to generate samples by solving an SDE, which involves simulating the dynamics of a stochastic process that gradually diffuses the data distribution towards a target noise distribution.

Song et al., 2021a stated that Markov chains represent approximated stochastic process trajectories that satisfy certain stochastic differential equations (SDE). In a score-based DPM, the model learns to estimate the score function, which is the gradient of the log-likelihood with respect to the input data. This score function can be used to generate samples by solving an SDE, which involves simulating the dynamics of a stochastic process that gradually diffuses the data distribution towards a target noise distribution.

The SDE for the diffusion process can be explained with the equation 2.35, where the

coefficient of diffusion, a and drift b is defined over W_t as a standard Brownian motion, $t \in [0, T]$, where T is some finite terminal time horizon. The stochastic calculus provides a stochastic process that allows terminal distribution to converge to standard normal distribution $\mathcal{N}(0, I)$. Therefore, for a given forward diffusion process, DPM models are optimized to reverse in-time trajectories of the forward diffusion process with reverse diffusions.

$$dX_t = b(X_t, t)dt + a(X_t, t)dW_t \tag{2.35}$$

It is to be noted that the "time" T mentioned in this context is a conceptual idea related to the progression of the process and is not related to the time axis of an audio signal or its STFT.

In such a scenario, the generation process involves sampling random noise from $\mathcal{N}(0, I)$. Then, a simple first-order Euler-Maruyama scheme [Kloeden and Platen, 1977a] is used to solve SDE for describing the dynamics of the reverse diffusions. If forward and reverse diffusion processes have close trajectories, the distribution of resulting samples will be very close to that of the data. The recent scoring function-based DPM model uses SDE instead of Markov chains. Scoring function-based DPMs allows forward diffusions to transform to any data distribution of $\mathcal{N}(\mu, \sum)$, and not necessarily $\mathcal{N}(0, I)$, where μ is mean, and $\sum$ is the diagonal covariance matrix.

$$dX_t = \frac{1}{2}\sum^{-1}(\mu - X_t)\beta_t d_t + \sqrt{\beta_t}dW_t \tag{2.36}$$

$$dX_t = \left(\frac{1}{2}\sum^{-1}(\mu - X_t) - \nabla logp_t(X_t)\right)\beta_t dt + \sqrt{\beta_t}d\tilde{W}_t \tag{2.37}$$

For n-dimensional stochastic process X_t, SDE is defined by Equation 2.36 for transforming any data to Gaussian noise over finite time horizon T and noise schedule β_t as a forward diffusion process. Thereafter, SDE for reverse diffusion is defined by Equation 2.37 for reverse time dynamics of stochastic process of diffusion obtained from Anderson, 1982. In Equation 2.37 $d\tilde{W}_t$ is reverse-time Brownian motion, and p_t is the probability density function of the random variable X_t, where $t \in [0, T]$ to solve SDE backwards in time. In the score-based model, [Song et al., 2020b] SDE Equation 2.37 can be assumed to be an ordinary differential Equation, and thus Equation 2.37 can be rewritten as stated in Equation 2.38, where $\nabla logp_t(X_t)$ represent an estimate of the gradient of the log density of noisy data as score function of distribution $P(X)$. The $P(X) = P(X_t|X_0)$, is the probability distribution conditioned on the initial distribution of $P(X_0)$.

$$dX_t = \left(\frac{1}{2}\sum^{-1}(\mu - X_t) - \nabla logp_t(X_t)\right)\beta_t dt \tag{2.38}$$

$$dX_t = \left(\frac{1}{2}\sum^{-1}(\mu - X_t) - s_\theta(X_t, \mu, t)\right)\beta_t dt \tag{2.39}$$

$$\mathcal{L}_{Fisher-divergence} = \mathbb{E}_{p(x)}[\|\nabla logp_t(X_t) - s_\theta(X_t, \mu, t)\|_2^2] \tag{2.40}$$

According to forward Kolmogorov equations, Equation 2.36 and Equation 2.38 present the evolution of probability density functions of an identical stochastic process. Therefore, score based DPM model learns to approximate neural network $s_\theta(X_t, \mu, t)$ such that s_θ is the approximate estimate of $\nabla logp_t(X_t)$, which reduces Equation 2.38 to Equation

2.39. During the training of the DPM model, Fisher divergence is used to minimize the loss criterion between the model output and noisy data distribution defined as in Equation 2.40. The Fisher divergence measures the similarity between the ground-truth data score and the score-based model by comparing their squared distance. However, it's impossible to directly compute the divergence because it requires access to the unknown data score. Score matching address this challenge by minimizing the Fisher divergence without knowledge of the ground-truth data score. Score-matching objectives can directly be estimated on a dataset and optimised it using stochastic gradient descent, similar to how likelihood-based models are trained (with known normalizing constants). By minimizing the score-matching objective, we can train the score-based model without relying on adversarial optimization. A more detailed description of the derivation of the DPM model used in the scope of this thesis work is available in [Sohl-Dickstein et al., 2015, Ho et al., 2020, Song et al., 2020a].

2.7 Speech enhancement for SV

While noise robustness is a general and hard problem for many speech-processing tasks, there are relatively few studies that focus on using speech enhancement for speaker verification tasks. One of the primary reasons is that instead of having multiple pre-processing steps to remove noise, it's simpler and more effective to train speaker verification systems using a large and diverse dataset that includes various noise types. When trained on such a dataset, speaker verification systems naturally become more robust to noisy environments. This approach is particularly useful for large neural networks with many parameters. Additionally, speech enhancement's primary goal is to reduce noise and improve speech quality rather than ensuring optimal performance in downstream tasks like speaker verification.

For these reasons, only a few studies have explored speech enhancement for speaker verification [Ortega-Garcia and González-Rodríguez, 1996, Plchot et al., 2016, Michelsanti and Tan, 2017, rahman Chowdhury et al., 2018]. The i-vector approach using a Denoising Autoencoder (DAE) to enhance the quality of noisy speech signals has also been employed [Ortega-Garcia and González-Rodríguez, 1996, Plchot et al., 2016]. The DAE is designed to minimize the L2 loss between the output of the model and the clean speech. During training, both noisy-clean and clean-clean pairs are needed to prevent the DAE from deteriorating the quality of the clean signal. These studies have reported improvements in noisy and mismatched conditions but only marginal gains in clean and matched conditions. This is because the objective of the DAE is to generate outputs that are similar to the inputs. [Michelsanti and Tan, 2017] utilizes conditional generative adversarial networks for speech enhancement, with the aim of preventing degradation caused by artefacts and distortions. In this method, the speaker verification system is trained individually using enhanced clean speech. In [Chang and Wang, 2017], a DNN network is trained for estimating IRM to attenuate background noise and subsequently enhance speech and is given as input to the i-vector-based system. Attention mechanisms have also been used with speech enhancement for speaker verification [Shi et al., 2020]. This is because a neural attention mechanism can allocate different weights to different input features. This can hence highlight the relevant information and reduce the interference caused by irrelevant information. [Shon et al., 2019] proposed a loss function for speech enhancement on the basis of speaker verification feedback. They integrated the speech enhancement and speaker verification module into a single framework.

The joint optimization of speech enhancement and speaker verification is another contribution that shows the usefulness of speech enhancement. Among them, some jointly optimized or integrated weighted prediction error (WPE) and some variants of beamforming using speaker embedding model for reducing the equal error rate [Yang and Chang, ,Taherian et al., 2019a,Mošner et al., 2018]. [Zhao et al., 2019] incorporated speech separation into an end-to-end speaker verification system, using deep learning for both modules. The approach involves combining a deep ResCNN-based speaker verification module with a CRN-based speech separation module to form a larger, more complex neural network, which is jointly optimized during training.

2.8 Conclusion

This chapter first gave an overview of the speaker verification system and the speech enhancement approaches. DNNs produce state-of-the-art speaker verification and speech enhancement performance compared to conventional systems. We also gave an overview of some of the conventional approaches in both domains. In the first part, we discussed the conventional speaker verification systems and the evolution of DNN. We particularly discussed two DNN-based speaker verification systems that we have used in the framework of this thesis work, namely x-vectors and ECAPA-TDNN. DNNs have significantly improved the performance of speaker verification systems, but they are still affected by noise interference in distant scenarios. We discussed the feature extraction process and its importance for the speaker verification system. We analyzed various factors that degraded the speaker verification system and pointed out that speech enhancement can be used to remove noisy interference. As the main aim of this thesis work is to develop speech enhancement algorithms for improving the performance of speaker verification in noisy-reverberant far-field scenarios, we analyzed various approaches that have been proposed over the years.

Chapter 3

Datasets and evaluation metrics

Contents

3.1 Introduction

Generating or collecting data and building the machine learning models are crucial steps in the project while evaluation metrics are equally significant in measuring the effectiveness of the machine learning models. Datasets and evaluation metrics are important for detecting and preventing biases in machine learning models. The bias may arise due to skewed data distribution, inadequate representation of certain classes, or other factors. Using appropriate evaluation metrics and diverse datasets can help detect and mitigate such biases, ensuring that the machine learning models are fair and unbiased. In this chapter, we give a detailed explanation of the datasets and the evaluation metrics used in this thesis work.

3.2 Datasets

Technological advancement and the growing dependability of technology have prompted the need for "good" quality data. British mathematician and data scientist Clive Humby

rightly said that "data is the new oil". But it is time-consuming and expensive to collect or record quality data. Besides, machines or computers are still dumb, and if the data we feed the machines with are of bad or low quality, we cannot expect the machines to give us "quality" output. As author Greg M Perry quoted, "Machines do not have intelligence. A computer blindly follows your instructions, step by step. If you do not give detailed instructions, the computer can do nothing". It is important to have the right kind of data to make machine learning algorithms work, as the system will do what it is supposed to do by learning from the data. Moreover, deep learning or machine learning algorithms are data-hungry, which means they require a large amount of data. In a nutshell, the success of a system depends on the "data" we feed it with. Thus, data preparation is a crucial aspect of developing a deep-learning-based system.

We have used different datasets to train and evaluate speech enhancement and speaker verification models. This is mainly because speech enhancement datasets are designed to improve the quality of speech signals by removing noise and distortions, making the speech more intelligible and easier to understand. These datasets typically include speech recordings with different types and levels of noise and interference. On the other hand, speaker verification datasets are used to verify the identity of a person from its speech signal. These datasets contain speech recordings from multiple speakers, with some speakers providing multiple recordings to allow for speaker verification tasks.

We give an overview of some publicly available datasets used widely for speaker verification which we have used in our experimentation, followed by an explanation of the data generation process we have applied for speech enhancement.

3.2.1 Speaker verification datasets and challenges

Speaker verification datasets typically include speech recordings from many speakers, with each speaker providing multiple recordings under different conditions, such as different environments, speech styles, and channel characteristics. The dataset should be diverse and balanced to ensure the model is not biased towards any particular speaker or group of speakers. The audio recordings in the dataset can be either text-dependent, where the speaker utters a particular phrase or a set of phrases, or text-independent, where the speaker speaks freely without any particular constraints. Our speaker verification system is text-independent.

When designing a speaker verification system, it's crucial to consider the target application and adjust the training and testing data accordingly. The data utilized in developing the speaker verification system can typically be split into three parts: training (on which the system is trained), development (to monitor performance and allow the potential adjustment of parameters), and evaluation/test set (final evaluation). The evaluation set can usually be divided into several benchmarks, focusing in detail on selected aspects of the evaluation (such as cross-language trials and noisy environment trials).

NIST SRE

The NIST Speaker Recognition Evaluation (SRE) is a series of evaluations conducted by the National Institute of Standards and Technology (NIST) since 1996 to assess the performance of speaker recognition systems [Doddington, 1998]. The NIST SRE dataset is a widely used benchmark for evaluating speaker verification systems. The dataset is designed to reflect real-world scenarios for speaker verification.

The dataset consists of speech recordings from several NIST SRE evaluations conducted since the beginning of the evaluation series. It includes speech samples from over 4,000 speakers recorded in English and Mandarin Chinese. The recordings were made in various environments, such as telephone and microphone channels, and under different recording conditions, such as clean and noisy environments.

The NIST SRE dataset is divided into two main subsets: core and extended. The core subset contains speech recordings from a large number of speakers, while the extended subset contains additional recordings of the same speakers in various conditions. Both subsets include enrollment and evaluation data. The evaluation data is further divided into two sets: the development set and the evaluation set. The development set is used for tuning the system parameters and selecting the best model, while the evaluation set measures the system's performance on unseen data. The evaluation set contains several different trials, such as same-speaker and different-speaker trials, to evaluate the performance of the speaker verification system in various scenarios.

The NIST SRE challenge has been organised every two years since its beginning in 1996 to evaluate text-dependent speaker verification systems. The challenge included new tasks and evaluation metrics every subsequent time. The challenges have attracted participation from research groups and companies from around the world and have contributed to significant advances in speaker recognition technology. In addition to the regular evaluations, NIST has conducted special challenges, such as the Deep Learning for Speaker Recognition Challenge in 2018, which focused on using DNN for speaker recognition.

SITW

The Speakers in the Wild (SITW) database [McLaren et al., 2016] is a large-scale, open-source dataset for evaluating speaker recognition systems. It was developed by the National Institute of Standards and Technology (NIST) and released in 2016. The SITW dataset has become a benchmark for evaluating speaker recognition systems in unconstrained environments. It has been used in several speaker recognition evaluations, including the NIST Speaker Recognition Evaluation (SRE) 2016 and the VoxSRC Speaker Recognition Challenge. The dataset has also been used in research on deep neural networks for speaker recognition and for developing unsupervised speaker recognition techniques.

The SITW dataset contains speech samples from a diverse set of speakers recorded in real-world environments, such as telephone calls, YouTube videos, and radio broadcasts. The dataset includes more than 2,000 hours of speech from over 6,000 speakers, covering 12 languages and dialects. These audio recordings do not contain any artificially added noise, reverberation or other artefacts. It is commonly used for evaluating speaker verification systems in unconstrained scenarios.

The SITW dataset also includes metadata for each speech sample, including speaker demographic information, speech activity annotations, and channel information. The metadata can be used to evaluate the performance of speaker recognition systems on different speaker and channel conditions. The sitw-core-core benchmark comprises audio files, each of them containing a continuous speech segment from a single speaker. The enrollment and test segments contain between 6 and 180 seconds of speech each. SITW also introduced other evaluation benchmarks for participants that are motivated to focus on speaker diarization. In sitw-assist-core/multi, the Person of Interest (POI) segment is partially in the enrollment part of the trial, and participants can use it to find the other

POI segments in the enrolment audio. In sitw-assist/core-multi, the information about multiple speakers in test audio is known.

FFSVC

The Far-Field Speaker Verification Challenge (FFSVC) dataset [Qin et al., 2020] is a publicly available benchmark dataset for far-field speaker verification research. It was released in 2020 by Amazon as part of the Interspeech 2020 Computational Paralinguistics Challenge.

The FFSVC dataset consists of recordings from over 6,000 speakers collected in various environments such as home, office, car, and public spaces. The recordings were captured using far-field microphones commonly found in smart speakers, phones, and other devices. The dataset includes both clean and noisy recordings, with various types of noise, such as music, traffic, and background conversations.

The FFSVC dataset is designed to be challenging, with high variability in speaker characteristics, recording conditions, and noise levels. The dataset is split into training, development, and evaluation sets. The training set contains about 3,500 speakers, while the development and evaluation set each contains about 1,200 speakers.

In the following sections, we present the datasets we have used for training and evaluation of the speaker verification models.

VoxCeleb

VoxCeleb is a widely used large-scale dataset for speaker identification and verification tasks [Nagrani et al., 2017]. However, since its release, VoxCeleb has been used for various tasks besides speaker identification and verification, such as speech separation, image synthesis, emotion recognition, and face recognition. The VoxCeleb dataset consists of over 1 million utterances from 6,112 speakers, primarily drawn from celebrities in public media. The dataset includes both audio recordings and metadata such as speaker identities, gender, and nationality. Each identity is referred to as the person of interest (POI). VoxCeleb was originally introduced in 2017 as VoxCeleb1 and was later expanded in 2018 with VoxCeleb2, which includes more diverse speakers and languages.

The VoxCeleb1 dataset contains 1,251 POIs with more than 150,000 utterances and 350 hours of speech. The audio signal was extracted from the segmented videos after the face tracking was used to separate the video segments for each of the POIs. In the speaker verification scenario, 1,211 speakers with 148,642 utterances were selected for the training set, and the remaining 40 speakers with 4,874 utterances were allocated into the test set. The test set was organised into $37,720$ pairs for speaker verification. The authors stated that the utterances in the VoxCeleb1 contain various noises appropriate to evaluate the noise robustness of models.

VoxCeleb2 [Chung et al., 2018] contains over 1 million utterances from 5,994 speakers, including individuals from different age ranges, ethnicities, and regions of the world. The dataset is divided into three subsets: development set, evaluation set, and test set. The development set and evaluation set each contain about 1,000 speakers with a total of around 500,000 utterances, while the test set contains the remaining speakers and utterances. In addition to the audio recordings, VoxCeleb2 also includes metadata such as speaker IDs, gender, and nationality. The dataset covers a wide range of languages: English, Spanish, German, French, Mandarin, and others.

An additional step for automatic duplicate removal is added, which allows curating the bigger version of VoxCeleb1. We train the speaker verification models on Voxeceleb datasets.

VOiCES

The Voices Obscured in Complex Environmental Settings (VOiCES) dataset [Nandwana et al., 2019] is a collection of speech data created by the Johns Hopkins University Human Language Technology Center of Excellence (HLTCOE) and was released in 2019. VOiCES is a publicly available dataset for promoting research in speech processing (automatic speech recognition, speaker verification, etc.), acoustic signal processing (source separation, general enhancement, etc.), and audio classification (speech/non-speech classification, etc.). The data was recorded using distant microphones in acoustically challenging environments in real rooms. Twelve microphones were set around the room to record the data. Around 120 hours of audio data were recorded using each microphone.

The VOiCES dataset includes over 1,500 hours of speech from 2,000 speakers in various acoustic environments, including noisy and reverberant rooms, vehicles, and outdoor settings. The speakers in the dataset come from diverse backgrounds, ages, and genders and speak in multiple languages, including English, Arabic, and Mandarin.

The speech data in VOiCES is recorded using a variety of microphone arrays, including single, stereo, and eight-channel arrays. The dataset also includes metadata such as speaker identities, location, and acoustic properties of the environment. In addition, the dataset includes artificially obscured speech data, in which speech is artificially mixed with environmental noise or distorted to simulate challenging real-world conditions.

The VOiCES dataset includes a challenge task, namely VOiCES from a distance challenge, in which participants are asked to transcribe speech in noisy and reverberant environments. The distance challenge aims to evaluate speech recognition systems' ability to accurately transcribe speech from far-field microphones, which are often used in real-world applications such as smart speakers and voice assistants. Far-field microphones are subject to a number of acoustic challenges, including environmental noise, reverberation, and speaker variability. The challenge has two tasks: far-field automatic speech recognition and speaker recognition. The challenge is designed to encourage research on speech recognition in challenging acoustic conditions and to help evaluate the performance of new algorithms and models.

The distance challenge in VOiCES includes short, medium, and long distances. The short-distance condition involves microphones placed $1-2$ meters from the speaker, while the medium and long-distance conditions involve microphones placed $3-5$ meters and $6-10$ meters away from the speaker, respectively. In addition, the challenge includes speech data from multiple acoustic environments, including noisy and reverberant rooms, vehicles, and outdoor settings.

To evaluate the speaker verification models, we have used the VOiCES dataset.

MultiSV

MultiSV corpus was released to foster multichannel text-independent speaker verification research [Mošner et al., 2022]. They focus on tackling the lack of multi-channel training data by utilizing data simulation on the clean speech of the VoxCeleb dataset [Nagrani et al., 2017]. The development and evaluation trials are modified from the VOiCES corpus to provide multi-channel trials.

The MultiSV is designed for text-independent speaker verification and provides a training set of simulated 4 microphone arrays with approximately 77 hours containing background noise and reverberation. MultiSV dataset can also be used for dereverberation, denoising and speech enhancement.

Training

Due to the high cost of collecting multi-channel audio data, the authors use simulated training data. Simulated data is commonly used in the literature of multi-channel speaker verification and in other speech processing fields where a clean reference is required. The data is based on Voxceleb2 dev, which is often used for training embedding extractors. Additionally, the evaluation data include speakers from various datasets such as LibriSpeech in Libri-adhoc40.

The recordings obtained from the development section of the Voxceleb2 dataset are available in different channels and have varying noise levels. However, since these recordings are intended to be used as a source of speech in simulated data, the signals are preferred to be clean. To achieve this, a pre-selection process is conducted by estimating the SNR and retaining recordings with SNRs above 20dB only.

The remaining recordings were then split into two parts: the training set and the cross-validation set. The training set consists of speech from 1,000 speakers with an equal distribution of male and female voices (500 each). For each speaker, a maximum of 50 utterances were included, resulting in 72.26 hours of audio. On the other hand, the cross-validation set includes 90 speakers (45 females and 45 males), each represented by a maximum of 35 utterances, resulting in 4.68 hours. The average duration of utterances in both parts is 6.2 seconds.

To simulate indoor conditions, a point source of speech and a point source of noise is used. The noise sources were selected to mimic common indoor noise and were taken from three sources: (i)Music (amounts to 66.3 hours) - from the Free Music Archive [Defferrard et al., 2016], excluding recordings that are also present in the music section of the MUSAN dataset [Snyder et al., 2015], as these are used in the evaluation data. (ii) MUSAN noises (amounts to 5.0 hours) - 80% of the "noise" section of the MUSAN dataset is included, excluding music and babble. (iii) Freesound.org and self-recorded noises (amounts to 20.1 hours) - which includes sounds from real fans, HVAC, shop, crowd, library, office, and dishes sounds.

To generate reverberant speech and noise, the Image Source Method is used to perform RIR generation. A box-shaped room is simulated using 4-microphone Uniform Linear Arrays with a length ranging from 10 cm to 2 m. The RT60 was randomly sampled from a uniform range of $[0.3, 0.9]$ seconds. The generated RIRs were then convolved with the speech and noise sources. Mixing SNRs were uniformly drawn from $[3, 20]$ dB.

Development And Evaluation

The development and evaluation trials of the MultiSV dataset are based on those defined for the VOiCES challenge, and the same utterances are used. The data for the challenge was obtained by re-transmitting speech from LibriSpeech along with various types of background noises, such as babble, television, music, or none (diffuse background) noises. The audio was simultaneously recorded by multiple studio-quality and lapel microphones. However, the VOiCES challenge used single-channel enrollment and test segments, while

the MultiSV dataset contains multi-channel test recordings. These recordings were obtained by grouping four far-field VOiCES recordings with the same content to form ad-hoc arrays or by simulating microphone arrays. To support various use cases, multiple conditions were devised based on the properties of the enrollment segments, including CE (clean enrollment), SRE (single-channel re-transmitted enrollment), MRE (multi-channel re-transmitted enrollment), and MRE hard (similar to MRE but with the noise source playing in the background).

MultiSV has created development and evaluation sets with various lists of verification trials. The re-transmitted development set (dev retr) was obtained by selecting groups of four trials from the original VOiCES development set. Each group of four trials shared the same enrollment recording and the same test utterance but was recorded by four different microphones. These groups were reduced to a single trial with multi-channel test audio. The re-transmitted evaluation set (eval retr v1) was created similarly to the original VOiCES evaluation set, but it included microphones that may have negative SNRs and be confusing for single-channel enhancement models. Thus, a second version (eval retr v2) was defined, which replaced the problematic microphones with those close to non-problematic ones. We used the second version (eval retr v2) for our experimentation.

The simulated sets of multi-channel test segments used the same source speech as the re-transmitted sets but with simulated 4-microphone arrays. Therefore, the resulting development (dev simu) and evaluation (eval simu) definitions were the same as the re-transmitted ones regarding speakers, utterances, and the number of trials, but they had simulated microphone arrays. We have used the MultiSV dataset to train the speaker verification.

3.2.2 Speech enhancement datasets

A large amount of training data is required to create a well-performing system. And most DNN-based speech enhancement systems need clean speech references for training. However, such data collection is highly expensive because it necessitates recordings under well-controlled circumstances and requires maintaining the quantity and variety of acoustic conditions, speaker characteristics, and room geometry. Researchers have attempted to train DNN-based speech enhancement systems without clean speech data to overcome these issues [Alamdari et al., 2021]. The training technique employed in [**?**] was first proposed in Noise2Noise [Lehtinen et al., 2018] and requires pairs of noisy signals comprised of the same speech data but distinct noise samples.

Speech enhancement datasets are required to be carefully designed to ensure they are diverse, high-quality, and large enough to be effective. In the following sections, we describe some widely used datasets for multi-channel speech enhancement.

DEMAND

The Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [Thiemann et al., 2013] is a commonly used benchmark dataset for multi-channel speech enhancement. It was created by the Audio and Acoustic Signal Processing Group at the University of Edinburgh and contains various noise types and SNRs, making it a useful resource for evaluating the performance of multichannel speech enhancement algorithms.

The DEMAND dataset contains recordings of speech and noise from various real-world environments, such as offices, cafes, and streets. The recordings were made using multiple

microphones, allowing multi-channel recordings to be created. The dataset contains 600 recordings, each consisting of four microphone channels and a mixture of speech and noise.

The DEMAND dataset consists of a diversity of noise types and environments included in the dataset. This can help ensure that models trained on the dataset are robust to various real-world scenarios.

However, like any dataset, the DEMAND dataset also has certain limitations. For example, the dataset only contains a limited number of speakers, which may limit the ability of models trained on the dataset to generalize to other speakers. Additionally, the dataset only contains a limited number of recordings, which may limit the ability of models trained on the dataset to generalize to other environments and noise types. Moreover, the dataset only includes recordings with a limited set of noise types and levels, which may not fully capture the diversity of noise conditions encountered in real-world scenarios. This may limit the ability of models trained on the DEMAND dataset to generalize to other noise types and levels. This is why we did not use this dataset in our experimentation.

CHiME-4

The CHiME-4 dataset [Vincent et al., 2016] is a widely used benchmark dataset for multi-channel speech enhancement research, which was developed as part of the CHiME challenge [Barker et al., 2013]. CHiME-4 revisits the CHiME-3 dataset and makes it more challenging by reducing the number of microphones. The CHiME-4 dataset is based on the Wall Street Journal corpus (WSJ), which is a widely used speech recognition benchmark dataset. It consists of sentences spoken by talkers situated in challenging noisy environments, such as cafes, restaurants, and street scenes. The dataset was recorded using a 6-channel tablet-based microphone array.

CHiME-4 features three tracks depending on the number of microphones available for testing: a 6-channel track, a 2-channel track, and a 1-channel track. The 6-channel track is equivalent to the CHiME-3 setup [Barker et al., 2016], where all microphones are used. The 2-channel track focuses on 2-channel scenarios. The microphone pairs are selected for each utterance in such a way that microphone failures do not arise. The main challenge in this track is to mitigate the performance degradation from 6 to 2 microphones. The 1-channel track uses a single channel. The microphone is selected for each utterance in such a way that microphone failures do not arise. This track is similar to conventional ASR tasks based on single-channel processing. Important techniques used in this context could be single-channel enhancement, data simulation, and acoustic modelling.

The challenge provided two types of data: 'Real data' - for which speech data is recorded in real noisy environments (on a bus, cafe, pedestrian area, and street junction) uttered by actual talkers. 'Simulated data' - for which noisy utterances have been generated by artificially mixing clean speech data with noisy backgrounds. The ultimate goal is to recognise the real data. The clean speech data are based on the original WSJ0 training data. These are used to generate the simulated data of the training set.

While the CHiME-4 dataset includes a range of noisy environments, the set of noise types used is still limited. This can result in models that are not able to generalize well to different noise types. As we needed a dataset with more noise types and to fit both speech enhancement and speaker verification requirements, we did not use CHiME-4 in our experimentation.

TIMIT

The TIMIT database [Fisher, 1986] is a widely used corpus of read speech designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of ASR systems. It contains recordings of 630 speakers. Also, the recordings include eight dialects of American English. Each speaker in the dataset reads 10 phonetically-rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic, and word transcriptions. It also includes a 16-bit, 16kHz speech waveform file for each phrase said. The TIMIT corpus transcriptions have been hand-verified. The dataset has also been used for speech enhancement tasks.

However, the TIMIT dataset was primarily designed for speech recognition tasks and did not include recordings with significant amounts of noise or reverberation, which are common in real-world scenarios. This can limit the ability of models trained on TIMIT to generalize to real-world conditions where the SNR ratio is low.

NOIZEUS

The NOIZEUS database [Hu and Loizou, 2007] is a noisy speech corpus designed to facilitate the comparison of speech enhancement algorithms among research groups. It includes 30 IEEE sentences [Rothauser, 1969] spoken by six different speakers (three male and three female), which have been corrupted by eight different types of real-world noise at various signal-to-noise ratios (0dB, 5dB, 10dB, and 15dB). The noise signals are taken from the AURORA database [Hirsch and Pearce, 2000] and include different types of noises, such as train noise, babble, car, exhibition hall, restaurant, street, airport and train station noise. The signals are sampled at 8 kHz and saved in wave format (16 bits PCM).

3.3 Generating simulated dataset

We generated a synthetic dataset named RoboVoices, as collecting multi-channel data is time-consuming and expensive. Designing such a dataset is necessary as training speech enhancement approaches require ground-truth knowledge about the target speech and, to some extent, the degradation. There is a lack of multi-channel training datasets for speaker verification, and the publicly available datasets don't provide any ground-truth speech enhancement knowledge, which is an important aspect. This kind of information is not available in the available corpora for far-field speaker verification. We need a dataset that can fit both the speaker verification and speech enhancement requirements. Therefore, we designed this comprehensive dataset mainly for carrying out the research experimentation in the framework of this thesis for training the multi-channel speech enhancement and the speaker verification models.

Before describing RoboVoices, we give an overview of the datasets used as raw-material for generating the synthetic dataset.

3.3.1 Musan

The Musan dataset [Snyder et al., 2015] is a collection of audio recordings that are commonly used for research and development of speech and audio processing algorithms,

particularly for speaker recognition and speaker verification systems. The dataset consists of over 8,000 in audio files that are divided into three categories: speech, music, and noise.

The speech category includes recordings of speakers from various languages reading sentences, passages, and numbers. The music category includes recordings of various genres of music, such as classical, pop, and jazz. The noise category includes recordings of various types of environmental and artificial noise, such as street, office, and car noise.

The MUSAN dataset is commonly used as a background noise source for speaker recognition and verification systems. The noise recordings are used to simulate real-world environments where speech is often mixed with background noise. The speech recordings are used to train and test speech recognition and speaker verification algorithms, while the music recordings are used to test the algorithms' ability to distinguish between speech and music.

Musan consists of approximately 109 hours of audio data available in OpenSLR[1]. Audio files are formatted into 16kHz WAV files. There are approximately 60.5 hours of speech data, out of which 20 hours and 21 minutes of data are read speech collected from Librivox[2]. The speech files are obtained from the Internet Archive and the Missouri Channel senate archives, which are entirely in English. The remaining 40 hours of recordings are collected from US government hearings, committees and debates. The recorded data for audio is a WAV file of an entire chapter of an audiobook per speaker.

For the noise category, data is recorded for around 6 hours (around 929 files). The noise data includes different types of noises, such as dial tones, fax machine noises, and more, as well as ambient sounds, such as car idling, thunder, wind, footsteps, paper rustling, rain, animal noises, etc. Although intelligible speech recordings are omitted, few recordings contain crowd noises with indistinct voices.

3.3.2 Freesound

Freesound[3] is a collaborative database of audio snippets, samples, recordings, and sound effects that are available to the public for free use. The platform was created by the Music Technology Group at the Universitat Pompeu Fabra in Barcelona, Spain, in 2005, and it has grown into a large community of audio enthusiasts and professionals who contribute to and use the platform.

The platform is designed to be a resource for anyone who needs audio content, such as sound designers, musicians, game developers, filmmakers, and other creatives. It features a search engine that allows users to find audio clips based on keywords, tags, and other criteria. Users can also browse through curated collections of audio clips organized by theme, genre, or other characteristics.

3.3.3 Generation process of RoboVoices

We generated RoboVoices, by simulating real room environments with clean speech, additive noise and reverberation. We generated 13,600 RIR configurations and split the dataset into train and evaluation. We designed a specific training set for multi-channel speech enhancement and an evaluation set with various RIR scenarios to be used for both

[1]https://openslr.org/
[2]https://librivox.org/
[3]https://freesound.org/

Table 3.1: Overview of the RoboVoices dataset.

	Training	***Evaluation***
No. of speech files	10000	1200
No. of noise files	3725	1000
No. of RIR rooms	10000	3600
No. of hours	25	2

speech enhancement and speaker verification. Table 3.1 gives a statistical overview of the RoboVoices dataset.

Speech

We have used the dry speech (raw) data from a clean subset of Librispeech [Panayotov et al., 2015] corpus as we need a clean or dry speech corpus to generate the data. Librispeech is approximately 1000 hours of 16kHz read English speech data collected as part of the Librivox project, prepared by Vassil Panayotov and Daniel Povey. We randomly selected around 10000 files from the clean subset of Librispeech and truncated them to 10 seconds duration for the training set of speech enhancement, contributing to 25 hours of speech data.

We use the Fabiole speech corpus [Ajili et al., 2016] to evaluate the speaker verification system. Fabiole is a French speech corpus comprising around 3000 excerpts spoken by 130 native French speakers. The speech data of the corpus was collected from different French radio and television shows. For creating each evaluation set, we have used 1200 speech files from Fabiole representing 2 hrs of evaluation material.

Noise

We prepared a noise dataset by collecting realistic noises from the Freesound platform[4] and dividing them into two sets: RoboVoices noise train set for training and RoboVoices noise eval set for evaluation. We have convolved the dry speech from Librispeech and RoboVoices noise train set with simulated RIR for training. The training set consists of 3725 clips of noises selected from different categories such as door, keyboard, office, phone, background noise inside the room, printer, fan, door knock, air conditioning, babble, environmental noise, etc. This diversity of noise categories in the training set can help the system to learn to recognize and remove different types of noise effectively.

We have also created two evaluation sets: RoboVoices eval1 and RoboVoices eval2. RoboVoices eval1 contains the dry speech from Fabiole and noise from the RoboVoices noise eval set convolved with simulated RIR for evaluation. RoboVoices eval2 contains the dry speech from Fabiole and noise from the MUSAN dataset collected from the OpenSLR platform[5] convolved with RIR for evaluation, including noise categories such as dial tones, raindrops, etc.

Having separate evaluation sets is important to evaluate the performance of the speech enhancement and speaker verification system on unseen data and under different noise conditions. This can help in understanding the limitations of the system and identifying areas for improvement.

[4] https://freesound.org/
[5] https://www.openslr.org/index.html

Reverberation and room impulse response (RIR)

The reverberation effect takes place when the speaker or other sound sources are far from the microphone.

To simulate room effects, we have generated an RIR corpus of 10000 rooms for training and 3600 for evaluation with pyroomacoustics toolbox [Scheibler et al., 2018]. For training, the room length was chosen between $[3-8]$ m, width was chosen between $[3-5]$ m, and the height was chosen between $[2-3]$ m. The absorption coefficient was drawn randomly such that the room's RT60 was between $[200-600]$ ms. The minimum distance between a source and the wall is 1.5 m and 1 m between the wall and the microphones. The RIR for the evaluation set was generated with the same room configuration as in the training set, but the absorption coefficient was selected to obtain an RT60 of 400 ms.

A script has been created to generate RIR for desired RT60 and the angle between the microphone and the source position. To simulate the reverberated clean speech and noise, we convolved dry noise and dry speech. Moreover, for data generation, we have re-sampled the noise signal to match the sampling rate of the clean speech. The SNR is set differently in training and evaluation.

Composition of the training and evaluation set

We first convolved the speech and noise with RIR from the same room. The noise segment of the desired duration is randomly selected from the noise samples of the RoboVoices noise corpus for the training set. We then added the convolved clean speech and convolved noise to obtain the mixture signal. For the training set, the SNR is drawn randomly with a uniform distribution between 0 dB and 10 dB. We select SNR at 5 dB, 10 dB, and 20 dB for the evaluation set. In total, we generated 10000 mixtures for training and 3600 mixtures for evaluation using the Fabiole dataset.

3.4 Evaluation

Performance evaluation of machine learning models is an integral part of the model development process, and so is the development of new algorithms, which requires relevant metrics. Using different evaluation metrics helps understand and assess a model's efficacy and performance, strength and weakness. Typically, there are two approaches for evaluating the performance of a speech enhancement model, namely, objective and subjective. Objective evaluation measures the quality of the output signal to the ground-truth isolated sources using a set of calculations to determine the quality of the separation. Subjective evaluation entails assigning ratings for the output of the source separation system by humans. Subjective evaluations are time-consuming and expensive. To comprehend and analyze the performance of each model and algorithm used in this thesis work, we used objective measures to evaluate the speech enhancement and speaker verification models.

We evaluated the speech enhancement performance using the metrics to estimate the distortion on the target signal and uncorrelated interference. We evaluated the speaker verification performance using the commonly used equal error rate (EER) metric.

3.4.1 Speech enhancement metrics

Measuring the performance of a speech enhancement approach is a challenging problem. Moreover, speech enhancement mainly focuses on two goals:

- to make the speech easily understandable by distinguishing sounds from the background noise.

- to make the speech more intelligible and perceivable by alleviating the distracting interference of noise.

Signal-to-distortion Ratio (SDR) and signal-to-interference Ratio (SIR) [Vincent et al., 2006] has been used widely to measure the performance of speech enhancement and are particularly well suited for studying the characteristics of various speech enhancement algorithms. They decompose the estimated signal into a target source (clean speech), interfering with distortions, artefacts, and noise.

An estimate of a source $\hat{s}_i$ is assumed actually to be composed of four separate components:

$$\hat{s}_i = s_{target} + e_{interf} + e_{noise} + e_{artif}$$

where s_{target} is the ground-truth, e_{interf} is the error term for interference, e_{noise} is the error term for noise, and e_{artif} is the error term for the artifacts.

Signal-to-interference ratio (SIR)

SIR estimates the residual interference or noise from other sources while performing separation. SIR can be defined as:

$$SIR := 10log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2} \tag{3.1}$$

Signal-to-artefacts ratio (SAR)

SAR estimates the 'musical noise,' i.e. isolated sounds in frequency and time. SAR can be defined as:

$$SAR := 10log_{10}\frac{||s_{target} + e_{interf} + e_{noise}||^2}{||e_{artif}||^2} \tag{3.2}$$

As SAR didn't bring much information in our experimentation, we have not reported the results with SAR.

Signal-to-distortion ratio (SDR)

SDR was proposed for evaluating the audio blind source separation in the BSS_eval toolkit [Févotte et al., 2005]. The toolkit suggested different versions for the metric that might factor out certain effects, such as different signal gains or time-invariant filtering, which are insignificant in evaluating distortion. SDR is usually considered an overall measure of how good an estimated source sounds compared to the original sound. SDR can be defined as below:

$$SDR := 10log_{10}\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2} \tag{3.3}$$

3.4.2 Speaker verification metric

The efficiency of a speaker verification system is reflected in the demonstration of their accuracy and effectiveness in the final decision-making stage. The process of recognizing a speaker involves comparing the extracted features of their speech signal to the stored features of the enrolled speaker. Based on the distance or similarity values between the two sets of features, a decision is made as to whether the two utterances belong to the same speaker. The determination of speaker identity relies on the calculation of a similarity score between the two specified feature sets. If the calculated score is higher than the pre-defined threshold, then the two utterances are assumed to be from the same speaker. If the similarity score is below the threshold, it is rejected as coming from a non-enrolled speaker. According to Oglesby, 1995, the key variable in speaker verification is the value of the threshold. However, setting the threshold too high increases false rejections, while setting it too low increases false acceptances.

If a mistake is made during speaker verification evaluation, it will either result in a false accept or a false reject. The false accept error essentially grants access to an imposter, whereas a false reject error prevents a genuine speaker from being heard when the speaker verification system authenticates speakers. The false accept rate is calculated as the number of false alarm mistakes from a given number of imposter attempts, while the false reject rate is calculated as the number of false reject errors from a given number of legitimate efforts. Therefore, evaluating the system using an equal error rate (EER) is a better evaluation method than other methods [Sztah'o et al., 2019], like using classification accuracy.

Equal error rate (EER)

Equal error rate (EER) is a common performance metric used in speaker verification and other biometric identification systems. The EER is the point at which the rates of false acceptance and false rejection are equal. In other words, it's the threshold where the probability of falsely accepting a different speaker as the same as the target speaker is equal to the probability of falsely rejecting the target speaker as a different speaker.

The EER provides an indication of the overall performance of the system in terms of its ability to verify the identity of speakers correctly. A lower EER indicates better performance because it means that the system is able to balance the false acceptance and false rejection rates more effectively. The EER is typically reported as a percentage value and is commonly used to compare the performance of different speaker verification systems.

3.4.3 Evaluation protocol

During the evaluation, we generated enhanced speech signals with different acoustic conditions, such as noisy (mixture of noise and reverberated clean speech), dry clean speech (non-reverberated clean speech), and reverb-target (reverberated clean speech).

We evaluate the speech enhancement models in terms of SDR and the SIR on the input mixture (to obtain a reference point) and the signals estimated with the different enhancement approaches. We evaluated the speech enhancement systems on the RoboVoices dataset.

We evaluate the speaker verification in terms of EER on the clean and dry speech (as a reference point), the input mixture, and the signals estimated with different speech

enhancement metrics. We evaluated speaker verification on RoboVoices, MultiSV and VOiCES datasets.

3.5 Conclusion

This chapter enlisted publicly available datasets for speaker verification and speech enhancement. First, we give an overview of some of the widely used datasets for speaker verification. Then we describe the datasets used in this thesis work for the training and evaluation of the speaker verification systems.

Publicly available multi-channel speech enhancement datasets are limited in terms of the range of acoustic conditions they cover for the training and evaluation of speech enhancement algorithms. For example, they may not include sufficient amounts of noisy or reverberant speech. Moreover, most of the datasets lack ground-truth knowledge about the target clean speech, which is required during the training. Also, we needed a dataset that could be used for both multi-channel speech enhancement and far-field speaker verification. Therefore, to address this issue, We designed a synthetic dataset which we introduced in this chapter.

We also briefly describe the evaluation protocol followed while evaluating the models. We considered several acoustic scenarios to assess the performance of all the models used in this thesis work. For comparison purposes, we also considered dry clean speech (non-reverberated speech), noisy (mixture of noise and reverberated clean speech), and reverb-Target (reverberated clean speech).

This chapter describes several evaluation metrics for speech enhancement and speaker verification models. Speech signals are subject to being affected by additive noise, reverberation, and other distortions. The evaluation metrics in speech enhancement assessed speech quality in terms of reducing the noisy distortions and reverberation from the speech signal. As speaker verification aims to authenticate the identity of a speaker, the output of the speaker verification system is a similarity score, and a pre-defined threshold determines whether the result is positive or negative. EER in the context of the speaker verification system, helps assess whether two utterances are from the same speaker. In this case, rather than using classification accuracy, an EER is a better way to evaluate the system.

Chapter 4

Multi-channel speech enhancement for far-field speaker verification

Contents

4.1 Introduction

The use of microphone arrays in devices we use every day enables the application of multi-channel speech enhancement techniques, which makes it feasible to use these mobile devices in environments with multiple sources of interference. This allows for using speech-based applications in everyday scenarios where we are frequently surrounded by different kinds of noises, such as environmental noise like car honking, noises coming from the air conditioner or fan, and overlapping speech. Moreover, even if we are speaking in a completely silent room, every time we speak, our speech gets reflected by the wall of the room or the objects present in the room. This creates reverberation. Both these kinds of phenomena distort the quality and intelligibility of the speech signal, making it difficult for applications like speaker verification to perform well, especially in a distant or far-field scenario.

The speech signals in far-field speaker verification face various challenges that degrade their quality, making it more difficult to recognize the speaker's voice accurately. When speech signals travel over long distances, they experience attenuation due to absorption, scattering, and diffraction effects. This attenuation is called long-range fading and can lead to a decrease in the SNR of the speech signal. As a result, the speech signal may become weaker and more difficult to recognize. In real-world scenarios, there are often various environmental noises, such as traffic, construction, and other sources of ambient noise. These noises can mask the speech signal and make it more difficult to recognize the speaker's voice. Moreover, in some scenarios, the acoustic environment can be complex and challenging for speaker verification systems to operate effectively. For example, in a crowded room, there may be multiple speakers and sources of background noise, making it difficult to distinguish the target speaker's voice from other sounds. The placement of microphones in a room can also impact the performance of speaker verification systems. If microphones are too far away from the speaker, the speech signal may be weakened by long-range fading, and background noise may become more dominant. On the other hand, if microphones are too close to the speaker, the speech signal may be distorted by proximity effects, such as popping and breathing sounds.

As we have explained in section 2.5, speech enhancement can improve the quality of degraded speech signals. Multi-channel approaches aim to obtain a clean signal from a noisy mixture recorded with multiple microphones. These approaches effectively take advantage of spatial information for better performance in challenging scenarios. The state-of-the-art multi-channel speech enhancement systems mostly rely on DNN to estimate either a multi-channel filter or the estimated speech directly [Heymann et al., 2015, Nugraha et al., 2016, Ribas et al., 2022]. Depending on the design of the DNN, the multi-channel filter may be linear or non-linear. In the case of a linear filter, the DNN applies a set of fixed coefficients to the input microphone signals. These coefficients can be pre-determined through various signal processing techniques or learned during training. The output of the filter is a linear combination of the input signals, which can be used to enhance the speech signal or reduce noise. In the non-linear case, the DNN is trained to learn a mapping from the microphone signals to the speech signal, taking into account the effects of noise and reverberation. This mapping is learned using a large dataset of speech and noise signals, where the speech signal is the target output, and the noise signal is the input. The DNN learns to extract the speech signal by modelling the statistical properties of the input signals and their relationship to the target output.

In this chapter, we describe the proposed multi-channel speech enhancement pre-

processing to speaker verification in adverse acoustic conditions where noise and room reverberation distorts the target speech signal. We present this work as a benchmark of multi-channel speech enhancement approaches for far-field speaker verification. We consider either filtering based on DNN or combining DNN and signal processing approaches. The DNN-based approach we use implements FaSNet, a state-of-the-art neural beamforming technique for speech enhancement. We use FaSNet to compute the time-frequency (T-F) masks that inherently consider phase information. The second approach integrates FaSNet with Rank-1 multi-channel Wiener filter (MWF) and weighted prediction error (WPE), a de-reverberation algorithm. This chapter also presents the impact of the FaSNet-based approach and FaSNet Rank-1 MWF WPE approach in different noisy and reverberated acoustic scenarios under various signal-to-noise (SNR).

4.2 Neural pre-processing approach

In this section, we explain the speech enhancement pipeline we developed. We use FaSNet, a neural beamforming technique, to separate noise and speech. We use this first speech and noise estimations to compute the speech and noise covariance matrices that are, in turn, used to compute the Rank-1 MWF [Souden et al., 2010, Wang et al., 2017]. We used the Speech Distortion Weighted (SDW) variant of the Rank-1 MWF to estimate the filter [Doclo and Moonen, 2002a]. The output of the Rank-1 SDW-MWF is further processed with WPE [Yoshioka and Nakatani, 2012] to attenuate the reverberation. The entire processing pipeline is represented in Figure 4.1.

4.2.1 FaSNet

We applied FaSNet [Luo et al., 2019], a filter-and-sum time-domain neural beamforming network incorporating a two-staged architecture. The first stage computes a beamforming filter for a randomly chosen reference channel, while the second stage uses the output filter from the first stage to estimate the beamforming filters for the remaining channels. The input for both stages includes the target channel to be beamformed and the normalized cross-correlation (NCC) output between channels as an inter-channel feature. The NCC is a measure of similarity between two signals, which in this case, are the signals recorded by the reference microphone and the other microphones in the array. The NCC is calculated at the frame level, which means that it is calculated for each frame of the input signal. By using the NCC as an inter-channel feature, FaSNet is able to incorporate information about the correlations between the different channels in the array. This helps the model estimate the beamforming filters more accurately, which in turn improves the quality of the enhanced signals. Both stages use the temporal convolutional networks (TCN) [Lea et al., 2016]), enabling the low latency processing of the FaSNet model.

The training objective of the FaSNet model is to use a loss criterion at the signal level based on the actual task to be solved. Specifically, the SI-SDR and SI-SNR loss functions are used together. The SI-SDR (source-to-distortion ratio) loss measures the similarity between the estimated source and the ground truth source, while the SI-SNR (source-to-noise ratio) loss measures the quality of the estimated source signal by comparing it to a reference noise signal.

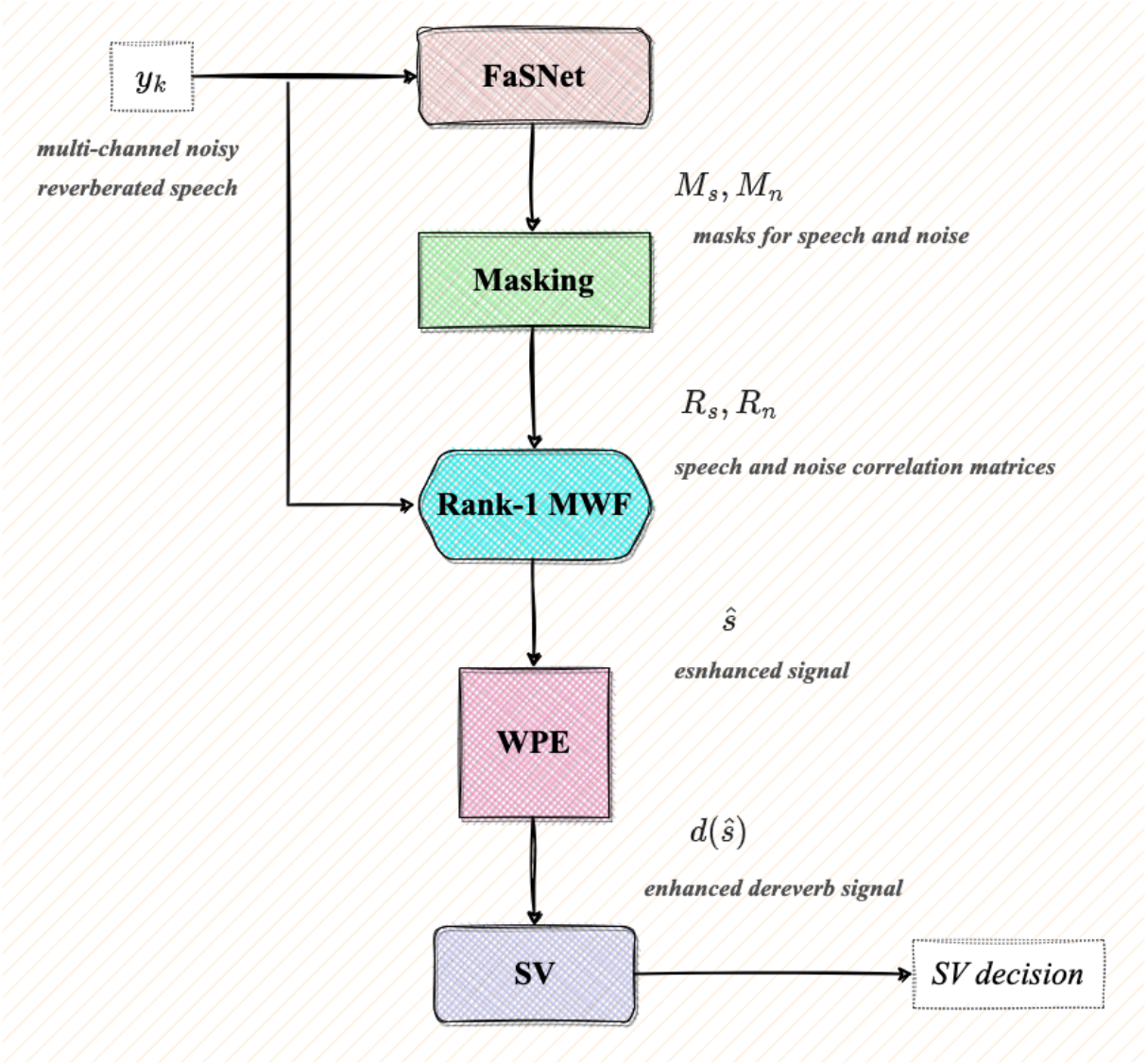

Figure 4.1: Graphical representation of the proposed pre-processing pipeline used in our experiments. We give a multi-channel noisy reverberated speech y_k as input. We used the output of FaSNet to estimate the masks for speech and noise M_s, M_n. These estimations are used to compute the covariance matrices for speech and noise R_s, R_n that are, in turn, used to compute the Rank-1 MWF. We apply WPE on the enhanced signal $\hat{s}$ to reduce reverberation. The enhanced dereverb signal $d(\hat{s})$ is given as input to the speaker verification.

4.2.2 Rank-1 multichannel Wiener filter

Over the years, linear filters, such as multi-channel Wiener filter [Doclo and Moonen, 2002b], Generalized Eigenvalue (GEV) [Warsitz and Haeb-Umbach, 2007], have been applied to reduce speech distortions caused by environmental or acoustic noise. In this thesis, we rely on noise reduction techniques based on MWF.

To provide an explicit tradeoff between the interference reduction and the signal distortion, we use the speech distortion weighted multichannel Wiener filter (SDW-MWF) described in section 2.5.2.

Under the assumption that there is only one source of speech, the correlation matrix

of speech $\mathbf{R}_s$ can be assumed to be a rank-1 matrix. Forcing this matrix to its Rank-1 approximation leads to the so-called Rank-1 version of the MWF filter.

The computation of MWF requires the estimation of the speech and noise correlation matrices. Until recently, this was done using a voice activity detection that is now widely replaced by T-F masks estimated with DNN [Furnon et al., 2020]. Most DNNs that are only used to estimate masks that take only single-channel input, but using multi-channel signals when available, have been shown to improve overall performance [Furnon et al., 2020, Perotin et al., 2018]. In this work, we use FaSNet to estimate the masks.

$$M_s(t,f) = \frac{|s(t,f)|}{|s(t,f)| + max(|s| + |n|, \varepsilon)} \tag{4.1}$$

$$M_n(t,f) = \frac{|n(t,f)|}{|s(t,f)| + max(|s| + |n|, \varepsilon)} \tag{4.2}$$

where, ε is 1×10^{-16}.

The predicted speech mask with DNN ($M_s(t,f)$) is applied to the mixture to obtain an estimation of the speech signal $\check{s}$:

$$\check{\mathbf{s}}(t,f) = M_s(t,f)\mathbf{y}(t,f) \tag{4.3}$$

The noise estimates $\check{n}$ can be estimated similarly with the noise mask predicted by the DNN ($M_n(t,f)$). These speech and noise estimates are used to compute the spatial correlation matrices $\mathbf{R}_s$ and $\mathbf{R}_n$ needed to derive the MWF using equation 2.34. The correlation matrices are obtained as follows:

$$\mathbf{R}_s = \frac{1}{T}\sum_{t=0}^{T-1} \check{\mathbf{s}}(t,f)\check{\mathbf{s}}(t,f)^H \tag{4.4}$$

4.2.3 Weighted Prediction Error

Dereverberation aims to reduce the reverberant components in the acquired speech signal while preserving the direct signal component [Naylor and Gaubitch, 2011]. Dereverberation using multiple microphones has received more interest, as the spatial filtering capability of the beamforming process can separate the reverberation part from the direct part.

Weighted prediction error (WPE) is one of the well-known dereverberation algorithms for mitigating performance degradation in speech-based applications. WPE is based on a time-varying power spectrum model, assuming that the desired source signal follows the pre-defined specific source priors such as Gaussian distribution [Naylor and Gaubitch, 2011]. WPE estimates the optimal filtering in the maximum likelihood sense using the observed reverberated speech signal. It is predicated on the idea that the desired signal follows a complex zero-mean Gaussian distribution with an unidentified time-varying variance.

We employed WPE [Yoshioka and Nakatani, 2012] as front-end processing combined with our multichannel speech enhancement. WPE aims to reduce the impacts of late reverberation by employing robust blind de-convolution with linear prediction. WPE can be used to alleviate speech recognition degradation performance, mostly in the case of a far-field scenario. The de-reverberated signal is obtained by subtracting the filtered signal from the observed signal denoted by:

$$\hat{x}(t) = x(t) - \Sigma_{k=1}^{N} \hat{w}(k) x(t-k) \tag{4.5}$$

where x_t is the reverberated signal at time t and $\hat{x}$ is the dereverberated signal using the WPE algorithm. $\hat{w}(k)$ denotes the k^{th} tap of the N-taps. WPE filter is $\mathbf{w} = [w_1, \ldots, w_N]^T$.

4.3 Loss function

This section describes the loss functions used in our experimentation. We used scale-invariant signal-to-distortion ratio (SI-SDR) and cross-entropy loss in our experimentation.

4.3.1 SI-SDR loss

The SI-SDR loss [Le Roux et al., 2019, Kolbæk et al., 2019] is the particular case of SDR (refer to section 3.4.1), allowing only scale invariance. Although, MSE loss is widely used between the predicted and target speech signals. However, the best speech quality may not always be achieved by minimizing the MSE. The scale-invariant signal-to-distortion ratio (SI-SDR) loss function has been shown to achieve better denoising performance [Kolbæk et al., 2019]. SI-SDR was introduced to evaluate the effectiveness of speech-processing algorithms. Unlike SDR, SI-SDR is scale-invariant to the processed signal, but finite-impulse response filters do not deform it. The SI-SDR can be expressed as follows;

$$SI\text{–}SDR = 10 \log_{10} \left(\frac{|| \propto s||^2)}{|| \propto s - \hat{s}||^2} \right) \tag{4.6}$$

where

$$\propto = argmin_{\propto} || \propto s - \hat{s}||^2 = s^T \hat{s} ||s||^2 \tag{4.7}$$

The scaling of the reference signal s ensures that the SI-SDR is invariant to the scale of $\hat{s}$. It can be seen from the equation 4.3.1 that SI-SDR is the SNR between the weighted clean speech signal and the residual noise.

The sample correlation between s and $\hat{s}$ is equivalent to maximizing the $\mathcal{L}_{SI-SDR}$ and measures the difference between the estimated signal and the ground-truth signal. SI-SDR loss function can be defined as follows;

$$L_{SI-SDR} = -SI\text{–}SDR \tag{4.8}$$

4.3.2 Cross-entropy loss

We used cross-entropy loss, which is traditionally used to train the speaker verification module. Cross-entropy computes the average difference between two probability distributions, the enhanced or predicted sample and the ground-truth sample. It's used to calculate a score that represents the typical discrepancy between predicted and actual data. It is better to have a lower score to improve the model's accuracy. The cross-entropy score ranges from 0 to 1, with 0 being the ideal number.

In our work, we used the cross-entropy loss function provided by the PyTorch library, which can be expressed as follows;

$$l(x,y) = L = \{l_1, ..., l_N\}^T, l_n = -w_{y_n} log \frac{exp(x_n, y_n)}{\sum_{c=1}^{C} exp(x_n, c)} .1 \tag{4.9}$$

where, x is the input, y is the target, w is the weight, C is the number of classes, and N spans the mini-batch dimension.

4.4 Experimental set-up

This section explains the experimentations of multi-channel speech enhancement and speaker verification systems. We also explain the baseline systems for multi-channel speech enhancement used in the experimentation.

4.4.1 Multi-channel speech enhancement

The speech and noise signals are sampled at 16 kHz in our experiments.

We provide multichannel speech signal as input to FaSNet with 4 ms window size and context size of 16 ms. We trained the FaSnet model with SDR loss and SI-SNR (scale-invariant source-to-noise ratio) loss [Le Roux et al., 2019]. We use the FaSNet implementation from the Asteroid Toolbox [Pariente et al., 2020]. We replaced the TCN blocks with a dual-path RNN (DPRNN) [Luo et al., 2020] in contrast to the original FaSNet architecture, where TCN is used to predict the beamformed filters. We employed the DPRNN with an encoder dimension of 50, a chunk size of 50, and a hop window of dimension 35 to implement the FaSNet-based neural beamforming. In the recurrency of DPRNN blocks with 4 layers, we use a hidden dimension of 128. We use the Adam optimizer to train the network for 200 epochs with a batch size of 4, a learning rate of 0.001 and a weight decay of $1e-5$.

We use the source-separated output of speech and noise from the FaSNet model to compute the target masks. The STFT is computed with a window length of 512 samples, a hop size of 256, and a Hann window. An SDW-MWF is estimated for each speech clip. We used WPE with the following parameters: 10 filter taps, a delay of 3 frames, 5 iterations of the WPE algorithm and alpha 0.9999. All metrics are presented with 95 % confidence interval (C.I) using bootstrap algorithm [Bisani and Ney, 2004].

We have used μ as a tradeoff factor between interference and distortion. In our experimentation, μ is varied from $0.1-0.9$ with Rank-1 MWF filter with IRM as shown in Table 4.1. After experimenting with various μ parameters in the Rank-1 filter computation, we observed that the μ value of 0.1 is the optimal hyperparameter for improving generalized performance across all the SNR conditions. We conducted a grid search strategy to find the optimal μ value, where experimentation suggested that the lower the μ value, the better speaker verification performance. We set the μ parameter of the SDW-MWF to 0.1 to limit the amount of distortion the filter introduces.

4.4.2 Baseline systems

In this section, we describe the baseline speech enhancement and speaker verification systems we have used in our experimentation.

Table 4.1: Experimental performance evaluation of μ parameters. The average confidence interval is 0.2.

μ	**Rank-1 MWF**		
	SAR	**SDR**	**SIR**
0.9	11.20	10.85	23.97
0.8	11.23	10.92	24.39
0.7	11.25	10.95	24.67
0.6	11.26	10.98	24.89
0.5	11.27	10.99	25.06
0.4	11.27	11.00	25.21
0.3	11.28	11.01	25.34
0.2	11.29	11.02	25.56
0.1	**11.30**	**11.05**	**25.65**

Speech enhancement

Taherian et al., 2019a examined several time-frequency masking-based beamforming methods for text-independent multi-channel speaker recognition in adverse acoustic conditions. The study evaluates different masking-based beamformers, including parameterized MWF, GEV beamformer, and MVDR beamformers. They also proposed to use Rank-1 approximation to estimate the speech covariance matrices, which are then used to estimate steering vectors in the MVDR beamformer. The GEV-BAN and Rank-1 approximated MVDR achieved superior performance than Rank-1 approximated parameterized MWF [Taherian et al., 2019a]. Therefore, we chose the best-performing GEV beamformer and Rank-1 approximated MVDR beamformer for our baseline system. Also, the network of both systems is closer to our proposed network. We didn't make any changes to the model architecture for comparative analysis.

The authors implemented BLSTM for the IRM estimation. The input feature is 129-dimensional log magnitude features, extracted using a frame length of 32 and a hop size of 8 ms. Global mean-variance normalization is performed on the features. The BLSTM network consists of 4 hidden layers, each with 300 units in each direction and an output layer with 129 sigmoidal units. The cost function is the MSE to estimate the T-F masks.

Speaker verification

The speaker verification used in this set of experiments is the standard Kaldi x-vector network introduced by Snyder et al., 2018. For the x-vector extractor, the training data is augmented with different sections from the Musan corpus (music, babble, noise, reverberation) [Snyder et al., 2015]. Then, we extracted MFCC features for the augmented data. The MFCC features are normalized by cepstral mean-variance normalization, and the silent frames are removed by VAD. The network is trained with 1 million augmented with all clean files from VoxCeleb1 [Nagrani et al., 2017] and VoxCeleb2 [Chung et al., 2018].

For test and enrollment files, the Fabiole corpus is used [Ajili et al., 2016]. In Fabiole, there are 6882 files from 130 speakers, out of which 3441 files are used for enrolment, and the remaining files are used for the test.

The PLDA (Probabilistic Linear Discriminant Analysis) classifier used for scoring is

trained on 200k x-vectors extracted from Voxceleb. Before training the PLDA, the x-vectors are centred, and their dimensionality is reduced to 128 with linear discriminant analysis.

We experimented with the speaker verification system for the match SNR and multi-SNR conditions. In speaker verification, the match SNR condition refers to a scenario where the SNR of the test data and the enrollment data are the same. This is done to ensure that any differences in performance between the two sets are not due to differences in the SNR. To achieve the match SNR condition, we select the same SNR value for both the enrollment and test data. This can be done by adding noise to the enrollment data to match the SNR level of the test data or by using denoising techniques to remove noise from the test data to match the SNR level of the enrollment data.

The advantage of using the match SNR condition is that it provides a more fair and accurate evaluation of the speaker verification system. If the SNR level of the test data is significantly different from the enrollment data, the performance of the system may be affected due to the differences in noise and signal characteristics. By ensuring that the SNR conditions of the enrollment and test data are the same, the match SNR condition helps to eliminate this potential bias and provides a more reliable evaluation of the system's performance. For the match SNR condition, we changed the enrollment data so that the test and enrollment set conditions are the same.

In speaker verification, the multi-SNR condition refers to a scenario where the SNR of the test data and the enrollment data varies across multiple levels. This is done to evaluate the performance of the speaker verification system under different SNR conditions. To achieve the multi-SNR condition, one approach is to randomly select SNR values for the enrollment data from a set of pre-defined SNR levels. During testing, the system is evaluated on test data at different SNR levels, which may or may not overlap with the SNR levels used for enrollment. This allows for a more comprehensive evaluation of the system's performance across a range of SNR conditions.

The advantage of using the multi-SNR condition is that it provides a more realistic evaluation of the system's performance. In real-world scenarios, the SNR of the speech signal can vary widely depending on the environment, microphone, and other factors. Evaluating the system under different SNR conditions helps to ensure that it is robust and effective in a variety of settings.

However, it is important to note that the use of different SNR levels during test and enrollment may also introduce additional variability and complexity to the evaluation process. Careful consideration must be given to the selection of SNR levels and the methods used to generate and manipulate the data to ensure a fair and accurate evaluation of the system's performance.

For the multi-SNR condition, the SNR for each enrollment file is drawn randomly in $[5, 10, 20]$ dB. We re-trained the PLDA scoring system for each condition using the Kaldi tool [Povey et al., 2011][6].

We also experimented with match pre-processing conditions, for which we applied the same pre-processing on the enrollment and the test data. In speaker verification, the match pre-processing condition refers to a scenario where the same pre-processing steps are applied to both the enrollment and test data. This is done to eliminate potential sources of variability in the data that could impact the performance of the speaker verification system.

The advantage of using the match pre-processing condition is that it provides a more

[6]https://github.com/kaldi-asr/kaldi

Table 4.2: Results on the RoboVoices eval1 and RoboVoices eval2 datasets using different pre-processing methods. RoboVoices eval1 consists of the dry clean speech from Fabiole and noise from Freesound. RoboVoices eval2 consists of the dry clean speech from Fabiole and noises from MUSAN. The average confidence interval is 0.1.

Noise type	**RoboVoices eval1**			**RoboVoices eval2**		
Pre-processing/SNR	**SDR**	**SIR**	**EER**	**SDR**	**SIR**	**EER**
Dry clean speech	—	—	14.9	—	—	14.9
Reverb. speech	—	—	20.6	—	—	20.6
Unprocessed (Noisy)	2.6	15.1	28.2	2.4	14.8	25.7
BLSTM Rank-1	5.4	20.7	26.8	5.1	20.7	23.1
BLSTM GEV-BAN	5.4	20.6	27.1	4.3	20.2	23.8
BLSTM MVDR Rank-1	5.8	20.1	27.0	4.9	20.5	23.7
FaSNet	5.3	20.5	38.7	4.7	18.3	32.6
FaSNet GEV-BAN	5.8	21,9	26.8	5.6	21.2	22.2
FaSNet Rank-1 MWF	6.1	21.0	24.9	5.9	21.5	21.9
FaSNet Rank-1 MWF WPE	**7.0**	**21.0**	**23.3**	**6.1**	**21.5**	**20.5**

fair and accurate evaluation of the system's performance. If different pre-processing steps are used for the enrollment and test data, the performance of the system may be affected due to differences in the pre-processed data. The match pre-processing condition helps to eliminate this potential bias and provides a more reliable evaluation of the system's performance.

4.4.3 Evaluation set-up

As we explained in Chapter 3, speech enhancement results were evaluated in terms of the source-to-distortion (SDR) ratio for estimating distortion on the target signal and the source-to-interference (SIR) ratio for estimating the relative importance of the estimated target speech compared to uncorrelated interference [Vincent et al., 2006][7]. The speaker verification system is evaluated using the EER.

We consider different conditions corresponding to different steps in the acoustic propagation process: Dry clean speech, Reverb-speech (reverberated speech) and Unprocessed (Noisy) (a mixture of reverberated noise and speech). We compute EER on dry clean speech and Reverb-speech (as a reference point), the Unprocessed (Noisy) and the signals estimated with different speech enhancement algorithms.

4.5 Results

This section reports the experimental results of the proposed approaches. To validate the proposed approaches, we compared them to the baseline BLSTM-based approaches [Taherian et al., 2019a]. The best results are highlighted in bold.

[7]SDR and SIR are computed with the mir_eval toolbox https://github.com/craffel/mir_eval

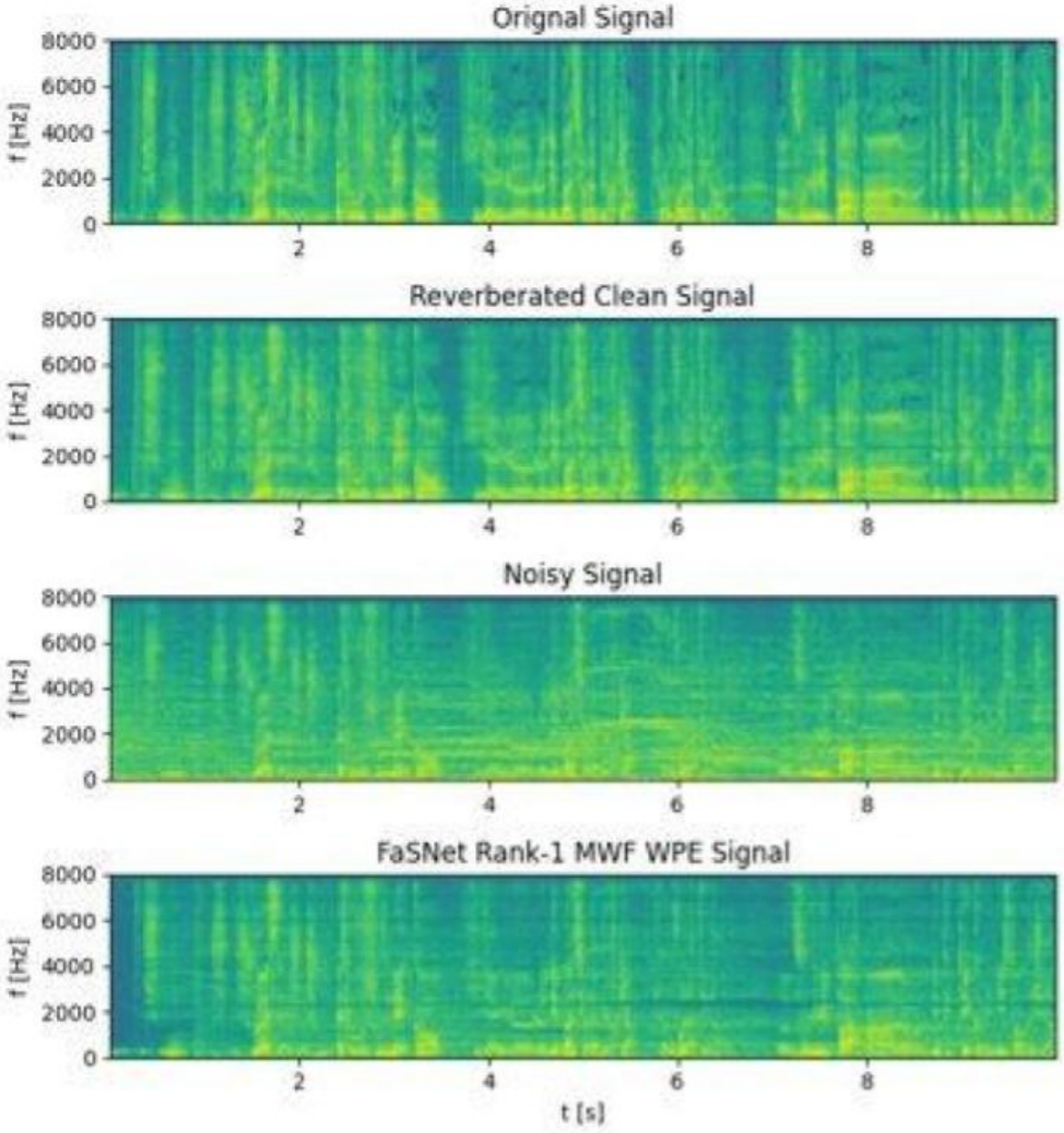

Figure 4.2: From top to bottom: spectrogram of original clean signal, the spectrogram of the reverberated clean signal, the spectrogram of the noisy (mixture) signal, the spectrogram of the enhanced signal with FaSNet Rank-1 MWF WPE at5 dB SNR.

Table 4.3: % EER on match pre-processing conditions on the RoboVoices eval1 dataset. We processed enrollment and test data by computing the SNR for 5 dB, 10 dB and 20 dB test data and averaged their EER. The average confidence interval is 0.1. Reverb. in the table refers to reverberated speech.

Test conditions	**Enrollment conditions**				
	Dry speech	Reverb speech	Unprocessed (Noisy)	MVDR Rank-1	FaSNet Rank-1 MWF WPE
Dry speech	**14.9**	15.4	16.7	16.1	15.7
Reverb. speech	20.6	**19.8**	20.5	20.4	20.1
Unprocessed (Noisy)	28.2	24.9	**23.8**	24.9	24,3
MVDR Rank-1	27.0	24.2	23.4	**21.3**	22.5
FaSNet Rank-1 MWF WPE	23.3	22.8	21.5	22.4	**19.2**

4.5.1 Comparison to baseline approaches

Table 4.2 presents the results of different state-of-the-art pre-processing approaches on the RoboVoices eval1 and RoboVoices eval2 datasets. We described both the sets in Chapter 4 section 3.3.3. We examine different masking-based beamforming approaches as pre-processing to far-field speaker verification under noisy and reverberated scenarios. The FaSNet based-approaches represent different pre-processing approaches we proposed in the framework of this work.

The performance is evaluated in terms of EER, SIR, and SDR, and the enrollment is always done using dry speech. The results show that noise and reverberation have a significant negative impact on speaker verification performance. The table also indicates that SDR is closely related to EER but not to SIR.

The table highlights that Rank-1-based beamforming techniques outperform FaSNet in terms of EER. This difference in performance could be due to the introduction of artefacts by FaSNet, as indicated by lower SDR values. Despite this, the integration of FaSNet, Rank-1 MWF, and WPE brings substantial improvement over all other techniques, indicating the benefits of using multiple pre-processing methods in combination.

The table also suggests that the RoboVoices eval1 dataset, which contains non-stationary noises, has a higher EER compared to the RoboVoices eval2 dataset. Nonetheless, FaSNet Rank-1 MWF WPE outperforms all baseline approaches based on BLSTM on both datasets. Overall, the results presented in Table A.1 demonstrate the effectiveness of using a combination of pre-processing techniques for far-field speaker verification under noisy and reverberated scenarios.

4.5.2 Impact of enrollment

We provide a set of results considering the impact of pre-processing on speaker verification depending on the speech enhancement approach, enrollment conditions, and its importance in improving speaker verification performance. Datasets and evaluation metrics used for experimentation are explained in chapter 3.

For the match pre-processing condition, we first computed an SNR for 5 dB test data, 10 dB test data and 20 dB test data, and then we averaged their EER. For enrollment,

Table 4.4: % EER on match SNR and multi-SNR conditions on the RoboVoices eval1 dataset. The average confidence interval is 0.1.

Test conditions	**Enrollment conditions**			
SNR (dB)	**Match SNR**			**Multi-SNR**
	5	10	20	
5	18.6	18.4	18.5	18.7
10	14.4	14.3	13.7	14.4
20	11.8	11.4	11.1	11.3

we created enrollment for each possible SNR.

Table 4.3 presents the EER averaged over SNR conditions for different pre-processing conditions of the speaker verification system, depending on pre-processing applied during the enrollment condition. Performance is reported on the RoboVoices eval1 dataset. For each enrollment condition, the EER values are shown for each test condition, namely dry, reverberated, noisy, and speech enhancement with BLSTM MVDR Rank-1 and FaSNet Rank-1 MWF WPE.

The results show that the choice of enrollment condition has a significant impact on the performance of the speaker verification system. Specifically, the EER values are generally lower when the enrollment condition matches the test condition. For example, for the dry test conditions, the lowest EER is obtained for the dry enrollment at 14.9%, while for noisy test conditions, the lowest EER is obtained at 23.8%. Performing the enrollment and test with matched acoustic conditions alleviates the effect of noise and reverberation. The results also show that FaSNet Rank-1 MWF WPE performs better than BLSTM MVDR Rank-1 in terms of EER in matched pre-processing conditions.

Pre-processing consistently improves speaker verification performance, but the effectiveness is more evident when the enrollment is done in matched pre-processing conditions (diagonal). We did not report the results with FaSNet as it still degrades the speaker verification performance due to the artefacts introduced by FaSNet. FaSNet Rank-1 MWF WPE obtained the best EER performance for a noisy and reverberated input.

Overall, the table demonstrates the importance of using matched pre-processing conditions in speaker verification, as well as the impact of different pre-processing techniques on system performance.

4.5.3 Impact of SNR

Table 4.4 presents the experimental results for speaker verification on the RoboVoices eval1 dataset. The speaker verification system is evaluated under two different conditions: match SNR and multi-SNR conditions. In the match SNR condition, both the enrollment and test data are pre-processed using the same SNR value (corresponding to only the diagonal values in the table). In the multi-SNR condition, the enrollment data is randomly chosen from the range of [5-20] dB SNR, while the test data is fixed at either 5 dB, 10 dB, or 20 dB SNR.

The rows of the table correspond to the SNR values used for testing, and the columns correspond to the SNR values used for enrollment. For example, the first row shows the results for testing with an SNR of 5 dB and enrollment with SNRs of $5, 10$, and 20 dB.

We observe that the SNR during enrollment has a minor impact. The impact is larger

Table 4.5: % EER on RoboVoices eval1 dataset using different pre-processing approaches.

Pre-proces./SNR	**5**	**10**	**20**
Unprocessed (Noisy)	34.4	28.0	22.2
BLSTM GEV-BAN	32.5	26.8	21.9
BLSTM Rank-1 MWF	32.2	26.5	21.6
BLSTM MVDR Rank-1	32.3	26.6	22.1
FaSNet	45.7	39.0	31.5
FaSNet GEV-BAN	32.0	26.3	22.0
FaSNet Rank-1 MWF	29.9	24.0	20.8
FaSNet Rank-1 MWF WPE	**27.1**	**23.2**	**19.7**

when the speaker verification operates on signals with a high SNR (above 10 dB); in this case, performing enrollment at low SNR (5 dB) decreases the speaker verification performance significantly. The results also suggest that doing enrollment with 20 dB SNR data always leads to good results, whatever the SNR level of the test data is. This indicates that it is better to train with clean data.

The multi-SNR condition enrollment results in only a slightly degraded performance compared to the match SNR condition. This indicates that the speaker verification system can operate without prior knowledge of the SNR conditions.

Overall, the experimental results suggest that the proposed speaker verification system is effective and robust to different SNR conditions. This could be because the SNR effect is compensated by the FaSNet Rank-1 MWF WPE pre-processing applied during both enrollment and the test.

4.5.4 Impact of pre-processing

Table 4.5 shows the performance of different pre-processing approaches for various SNR conditions. Performance is presented on the RoboVoices eval1 dataset. The first row of the table shows the EER for the "Noisy" condition, i.e., when no pre-processing is applied to the dataset. The following rows show the EER for different pre-processing approaches, such as BLSTM GEV-BAN, BLSTM Rank-1 MWF, BLSTM MVDR Rank-1, FaSNet, FaSNet GEV-BAN, FaSNet Rank-1 MWF, and FaSNet Rank-1 MWF WPE. For each SNR, the best results are displayed in bold (the approach providing the best results is mentioned in the next paragraph).

For each SNR condition, we observe an absolute reduction of EER for all the pre-processing approaches compared to an unprocessed input signal. This was expected for low SNR conditions but may be less obvious for high SNR conditions. The results show that FaSNet Rank-1 MWF WPE performs the best for all SNR values and outperforms other pre-processing approaches in terms of EER. With 7% EER absolute reduction at 5 dB, FaSNet Rank-1 MWF WPE shows the most robustness to low SNR conditions. Thus, supporting the argument that Rank-1 MWF is robust to low SNR scenarios.

4.5.5 Impact of utterance lengths

We experimented on the problem of short utterances as shown in Table 4.6. The problem of short utterances is well-known by the speaker recognition community. Short utterances

Table 4.6: % EER on different utterance lengths and SNR on RoboVoices eval1 dataset. Performance is averaged over SNR conditions for utterance lengths. The average confidence interval is 0.1.

Testing environment	**EER**	**EER**
Utterance length	**Below 4 secs**	**Above 4 secs**
Dry clean speech	18.8	5.6
Reverberated clean speech	21.1	7.5
Unprocessed (Noisy)	27.8	9.5
BLSTM Rank-1	27.4	9.2
BLSTM MVDR Rank-1	27.5	9.4
FaSNet	28.1	10.3
FaSNet Rank-1 WPE	27.5	9.0

are used to evaluate trials, which severely degrades the performance of the speaker verification system. However, the literature does not provide any standard definition of a short utterance. But, audios with shorter utterances containing less than 10 seconds of speech reported a severe degradation in quality by the speaker verification system [Mandasari et al., 2011, Kanagasundaram et al., 2016].

Table 4.6 shows the results of our experiments on utterance lengths below 4 seconds and above 4 seconds on the RoboVoices eval1 dataset. The SNR conditions are averaged to compute the EER for both utterance lengths. The baseline BLSTM-based and FaSNet-based models are trained using the same data for the proposed approaches. It is evident from the table that utterances below 4 seconds severely affected the speaker verification system. The EER for utterances below 4 seconds without pre-processing is almost 28%. Moreover, the baseline techniques don't seem to improve the performance either. In fact, FaSNet degrades the error rate compared to other pre-processing approaches mainly due to the artefacts FaSNet introduced during training. FaSNet Rank-1 MWF WPE manages to improve the performance by a margin. Nevertheless, it is observed from the results that short utterance lengths indeed degrade the signal quality of the speaker verification system.

4.5.6 Validation on a recorded unseen dataset

Table 4.7 presents the performance with various distractor noise conditions on the VOiCES evaluation dataset. Among 11 microphone positions in the Eval set, we selected 3 representative positions: closest to the speaker (microphone 2), mid-distance (microphone 9), and farthest (microphone 4). We selected the microphone closest to the speaker as a reference microphone for the FaSNet approach.

As expected, the condition with no noise distractor (clean in Table 4.7) resulted in the best performance for all the approaches. FaSNet Rank-1 MWF WPE still improves, probably because of reverberation. FaSNet is better than BLSTM Rank-1 for babble noise but not for TV and or Music noises. With an EER of 9.2% without any pre-processing, Babble seems to be the most challenging condition due to overlapping interference with the target speech. However, using FaSNet Rank-1 MWF WPE, the EER on Babble reduces to 6.3%. FaSNet Rank-1 MWF WPE achieves the best performance in all the noise conditions demonstrating the efficacy of our approach even though the model was trained

Table 4.7: % EER on different noise conditions of the VOiCES Eval dataset. The confidence interval is 0.2.

SV pre-processing	**Noise conditions**			
	Clean	Babble	TV	Music
Noisy	4.4	9.2	7.9	8.4
BLSTM Rank-1	4.4	8.1	7.1	7.3
BLSTM MVDR Rank-1	4.3	7.3	6.5	6.9
FaSNet	4.4	7.8	7.4	7.9
FaSNet Rank-1 MWF	4.5	7.1	6.8	7.1
FaSNet Rank-1 MWF WPE	**4.0**	**6.3**	**6.0**	**6.4**

on synthetic data generated for generic, possibly mismatched, and spatial scenarios. Notably, our approach generalizes to unseen noise, such as Babble. Designing a training set with matched conditions could help improve the performance of the VOiCES dataset.

4.6 Conclusion

Speech signals are contaminated with different kinds of distortions. These distortions can come from different sources, such as environmental noise and also due to the acoustic properties of the environment, as in the case of reverberation. They can severely degrade the speaker verification performance. In this chapter, we presented a benchmark of multi-channel speech enhancement for far-field speaker verification in noisy and reverberant environments. We experimented with DNN and a combination of DNN with signal processing methods as a front-end to the state-of-the-art x-vector-based speaker verification system. We proposed a multi-channel pre-processing pipeline by combining a time-domain neural beamformer (FaSNet), multi-channel Wiener filter (MWF), and weighted prediction error (WPE).

We analyzed the impact of the aforementioned approaches in realistic scenarios involving additive noise and reverberation on speaker verification performance. The integrated DNN and signal processing approach shows more robustness to low SNR scenarios. Experiments on synthetic and VOiCES datasets show that combining DNN with signal processing improves speaker verification performance. Primary efforts have been focused on the front-end multi-channel processing and the acoustic conditions during the enrollment phase.

We experimented by (i) implementing DNN-based FaSNet and observed that FaSNet is good at improving the performance of speech enhancement compared to baseline approaches. However, due to the artefacts introduced by FaSNet, as observed with low SDR values, FaSNet degrades the speaker verification performance. Secondly, combining the FaSNet with the Rank-1 MWF and WPE improved the performance across the noisy datasets. This suggests that usage of Rank-1 MWF with FaSNet and WPE assists in overcoming artefacts introduced by the FaSNet system.

We compared the integrated approach FaSNet Rank-1 MWF WPE to the existing state-of-the-art pre-processing approaches. We observed that the proposed approach brings substantial improvements in terms of EER over the baseline pre-processing approaches. WPE aids in dereverberation and consistently improves speaker verification performance. We evaluated the speech enhancement performance using the SDR and

SIR metrics to estimate the distortions on the target signal and the interference ratio compared to uncorrelated interferences. We experimented with different noise types to evaluate our system's robustness.

We examined the importance of pre-processing the enrollment data. The mismatch between enrollment and test conditions often deteriorates the performance of speaker verification. Real-world applications frequently control the enrollment condition as it is easier to control the enrollment conditions than the test conditions. On the other hand, the test condition is often considerably more uncontrollable and unpredictable. It varies from one test to another and can be significantly very different from the enrollment condition. We studied the impact of applying a matched processing on both the enrollment and test data. We also investigated to which extent the acoustic conditions during the enrollment phase should match the conditions during the test phase.

There is no proper description in the literature about the length of utterance for obtaining the best performance for a speaker verification system. However, often utterances shorter than 10 seconds leads to severe degradation. Therefore, we carried out the experiments on utterance lengths which are less than 4 seconds and vice-versa. We observed that all the systems achieved a higher equal error rate on the shorter utterance lengths, which proved that the longer the utterance length, the better the speaker verification performance.

The proposed FaSNet Rank-1 MWF WPE approach also leads to the best performance across the noise conditions on the VOiCES dataset even though the model was trained on synthetic data, showing generalization to the unseen real recorded data.

Chapter 5

Diffusion probabilistic models for multi-channel speech enhancement

Contents

5.1 Introduction

Generative models are one of the seminal tasks in comprehending the distribution of natural data. Diffusion probabilistic models (DPM), a new class of generative models first introduced by Sohl-Dickstein et al., 2015 in 2015, have achieved state-of-the-art performance in image [Ho et al., 2020] and text-to-speech processing [Popov et al., 2021, Jeong et al., 2021] as well as in speech enhancement [Lu et al., 2021, Lu et al., 2022]. The fundamental idea behind DPMs is that if we can create a learning model that can recognize the systematic loss of information caused by noise, we should be able to reverse the process and recover the information from the noise. However, DPMs are parameterized Markov chains designed to simulate a sequence of noise distributions in a Markov Chain rather than learning the actual distribution. It "decodes" the data by undoing/denoising

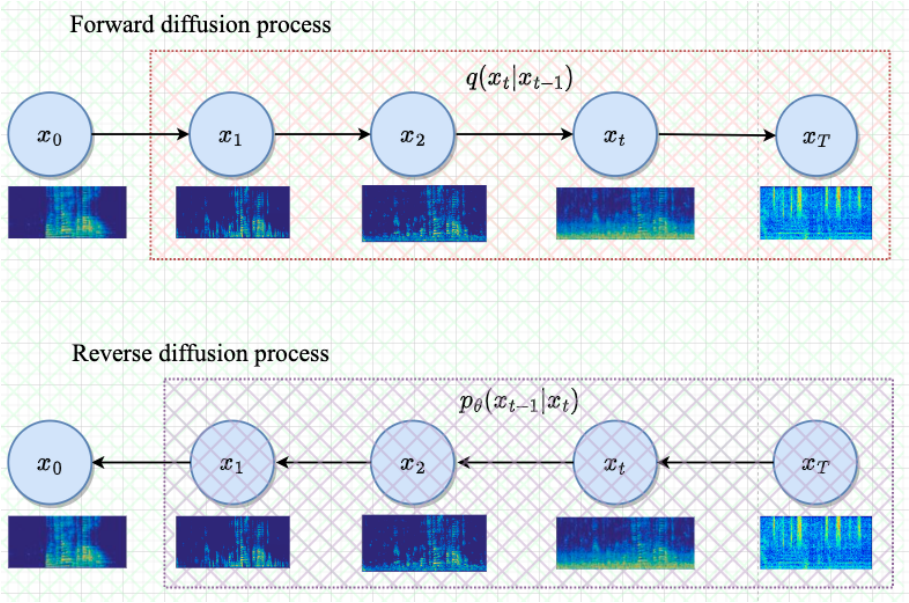

Figure 5.1: Graphical model of the forward and reverse diffusion processes. Each of the reverse conditionals $p_\theta(x_{t-1}|x_t)$ are structurally Gaussian and responsible for learning to revert each corresponding step in the forward process, i.e., $q_(x_t|x_{t-1})$. The mean and covariance of these reverse conditionals are neural networks with parameters θ and shared over timesteps.

it in a hierarchical manner. In a nutshell, diffusion models first describe a procedure for gradually turning data into noise and then train a neural network that learns to invert this procedure step-by-step. Each step consists of taking a noisy input and making it slightly less noisy by filling in some of the information obscured by the noise.

The key idea behind diffusion models is to model the distribution of the data by transforming a simple distribution, such as a Gaussian distribution, to a complex data distribution through a series of reversible transformations. Diffusion probabilistic models consist of two processes, namely, (i) a forward diffusion process and (ii) a reverse diffusion process. Figure 5.1 illustrates the forward and reverse diffusion process.

The forward process of diffusion models involves adding noise to a given data point. The noise is added in a probabilistic manner, typically using Gaussian noise. At each step, the level of noise is gradually increased, and the data point becomes more and more degraded. The goal of the forward process is to generate a sequence of noisy versions of the original data point, which can then be used for training the model. After training, we can use the diffusion model by applying the learned denoising process to generate data by simply passing randomly sampled noise. To simplify, by corrupting the data with added Gaussian noise in the forward process of the Markov chain in denoising the DPM-based approach, we meant that at each terminal step in the chain, we are simply sampling from a Gaussian distribution whose mean is the previous value in the chain [Ho et al., 2020].

The reverse process of diffusion models involves recovering the original data point from a sequence of noisy versions generated by the forward process. The reverse process is accomplished using a neural network to estimate the original data point at each step, starting from the most degraded version and working backwards towards the original. This is done by applying a sequence of transformations that reverse the noise-adding process at each step, effectively "undoing" the effects of the noise.

The denoising DPM involves training a Markov chain for both the forward and reverse diffusion processes in order to remove noise from a given input signal. However, this approach can lead to slower inference speed due to the need to perform multiple forward and reverse passes through the Markov chain. Moreover, in realistic conditions, the noise characteristics are usually non-Gaussian, which violates the model assumption when directly combining the noisy speech signal in the sampling process. Recently, few studies [Serrà et al., 2022, Welker et al., 2022] have explored scoring-based diffusion models for speech enhancement by implementing stochastic differential equations (SDE) instead of Markov chains.

The aforementioned papers for DPM-based speech enhancement are either based on the Markov chain or focus on single-channel speech enhancement.

The main focus of this chapter is the implementation and utilization of scoring-based DPM for multi-channel speech enhancement as a front-end processing for far-field speaker verification. We implemented a score-based DPM model to learn the time reversal using a stochastic differential equation (SDE) for sample generation and gradually diffusing the data distribution towards a particular noise distribution.

5.2 DPM-based multi-channel speech enhancement

This section gives a detailed theoretical review of the proposed frameworks for scoring-based DPM as a front-end processing approach we developed for a robust noise-aware speaker verification system under far-field noisy and reverberant scenarios. The scoring refers to the gradients of the log probability density of the noise [Song et al., 2021b]. The score-based generative model relies on computing gradients of the log probability density of noise with respect to a large number of data distributions that have been perturbed with noise. The framework of SDE introduces continuous-time dynamics to the diffusion process, where the level of noise is varied over time by solving an SDE. This helps us generalise conventional DPMs to reconstruct data from noise with different parameters, such as noise level and spatial correlations and allows us to make this reconstruction flexible by explicitly controlling the trade-off between sound quality and inference speed. Instead of perturbing data with a finite number of noise distributions, the score-based diffusion process is modelled by SDE that does not depend on the data and has no trainable parameters. The usage of SDE provides an easy-to-use framework for training DPM [Song et al., 2021b] but also controls the selection of the number of reverse diffusion steps for multi-channel speech enhancement.

We developed several multi-channel pre-processing techniques using DPM for speaker verification, which are explained in detail in the following sections. Our main contributions are as follows:

i. GradSE is a two-stage speech enhancement module for speaker verification.

ii. Diff-Estimator estimates the time-frequency masks using DPM.

iii. Diff-TasNet is a time-domain speech enhancement technique for speaker verification that replaces the diffusion model in U-net with a Conv-TasNet-based neural network architecture.

iv. Diff-Filter is used to compute the multi-channel filtered estimate of the clean speech signal in the time domain using DPM.

5.2.1 GradSE

We proposed to implement a speaker verification-dedicated multi-channel speech enhancement front-end named GradSE. The speech enhancement module has been named GradSE, as the primary function of the neural network used in the module is to compute the gradient of the log probability density of noise.

GradSE comprises a conditioning network and a DPM-based decoder network. The conditioning network defines conditional noise distribution by sampling from terminal distribution conditioned on the noisy signal in the diffusion process. In the context of diffusion models, the terminal distribution is the distribution of the data after it has been transformed by the diffusion process. The DPM-based decoder network is used for estimating gradients of the log density of noise.

We give the noisy multi-channel Mel spectrogram to the conditioning network, which outputs μ, the mean of the noise distribution, allowing for conditional generative modelling. We then give the clean speech Mel spectrogram and μ to the DPM-based decoder to perform the forward and reverse diffusion processes.

In GradSE's forward diffusion, the clean speech Mel spectrogram distribution is transformed into a terminal distribution, which is defined by conditioning network's output μ as $\mathcal{N}(\mu, I)$, where μ is the mean and I is the unit variance. After obtaining μ, it is parameterized with a latent variable sampled from a Gaussian distribution, which creates a conditional Gaussian distribution that depends on μ. This allows the neural network to learn how to deconstruct the clean speech Mel spectrogram into the noise distribution. By minimizing the scoring function in the forward diffusion process, the neural network learns to generate high-quality speech samples.

The terminal noise distribution of the forward diffusion process is parameterized by μ, allowing the decoder network to learn the trajectories of the forward diffusion process through the scoring function. The forward diffusion process slowly deconstructs the clean speech Mel spectrogram by adding noise to the data at each step of the diffusion. The forward diffusion process can be explained mathematically as below:

$$x_0 \sim \mathcal{P}_{data} \Longrightarrow x_T = \tau(x_0) \sim \mathcal{N}(\mu, I) \tag{5.1}$$

where $\mathcal{P}_{data}$ is data distribution of clean speech Mel spectrogram and in the forward diffusion process, which is referred by τ to slowly deconstruct x_0 by adding noise to the simple distribution defined by $\mathcal{N}(\mu, I)$ where, $\mathcal{N}(\mu, I)$ represents the distribution of noise which is added at each step of diffusion.

As explained in section **??**, stochastic processes, such as diffusion processes, in particular, are solutions of SDEs. In the forward process, SDE is used for adding noise perturbation to the data distribution. SDE perturbs data with Gaussian noise of mean zero and exponentially growing variance, which is analogous to perturbing data.

GradSE formulates a reverse diffusion process by a reverse SDE using the first-order Euler scheme [Kloeden and Platen, 1977b], which is then matched to the forward diffusion process in reverse time order. To achieve this, the neural network predicts the gradients of the log density of noise, which allows the forward diffusion to reconstruct the clean speech Mel spectrogram in reverse-time order.

In each step of reverse diffusion, the reverse trajectories of the forward diffusion are defined by SDE with an estimated scoring function from the decoder network. This iterative process transforms the conditioning network's output μ into the Mel spectrogram of the clean speech x_0. The reverse diffusion process is explained as given below;

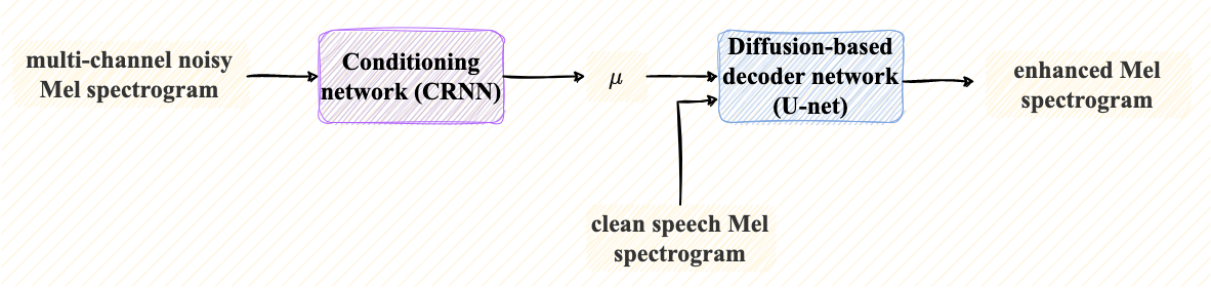

Figure 5.2: GradSE model in the training phase, which is composed of a conditioning network (CRNN) and a DPM-based decoder network (U-net). We give noisy multi-channel Mel spectrogram as input to the conditioning network from c number of channels with k number of Mel spectrogram frames. The output of the conditioning network, μ, represents the noise distribution. We give μ and clean speech Mel spectrogram as input to the diffusion-based decoder.

$$x_T \sim \mathcal{N}(\mu, I) \Longrightarrow \tau^{-1}(x_T) \sim \mathcal{P}_{data} \tag{5.2}$$

where, x_T denotes the noise sampled from terminal noise distribution defined by $\mathcal{N}(\mu, I)$ with μ as encoder output. τ^{-1} denotes the reverse diffusion process to construct the Mel spectrogram of the clean speech from data distribution, $\mathcal{P}_{data}$.

Training

As speaker verification takes Mel spectrogram as input, we give noisy multi-channel Mel spectrogram $y^{c,k,f}$ as input to the conditioning network, where c is the number of channels (microphones), k is the number of Mel spectrogram frames, and f is the dimension of the Mel spectrogram frames. Next, we use the conditioning network to compute μ^k, which is then used to estimate the noise distribution $\mathcal{N}(\mu^k, I)$. For simplicity of notation, we denoted the conditioning network's output, μ^k, as μ. Finally, we give the noise distribution of the conditioning network to the DPM-based decoder, which is then used to perform the forward diffusion process. During the forward diffusion process, the decoder learns the trajectories of the gradient of log probability to deconstruct the data into isotropic Gaussian noise distribution. This learned information is used in predicting reverse trajectories to remove the noise iteratively from sampled noise from noise distribution parameterized with encoder output μ. In the forward diffusion process, the decoder network iteratively deconstructs the Mel spectrogram of clean speech $x^k = x_0$ to noise x_T, where T is the terminal time horizon shown in Figure 5.2.

Loss function

We use two-loss criteria, mean square error (MSE) and diffusion loss. We applied MSE loss on the conditioning network's output with reference to the Mel spectrogram of clean speech. It is easier for decoding if we start from noise, which is already close to the Mel spectrogram of the clean speech (as MSE loss represents the latent log-likelihood estimate), to train GradSE. One of the objectives of using MSE loss is to minimize the distance between the distribution of the conditioning network's output μ and the target Mel spectrogram. The conditioning network is used to estimate the clean speech under MSE loss. The residual noise is computed by sampling from noise distribution, where the

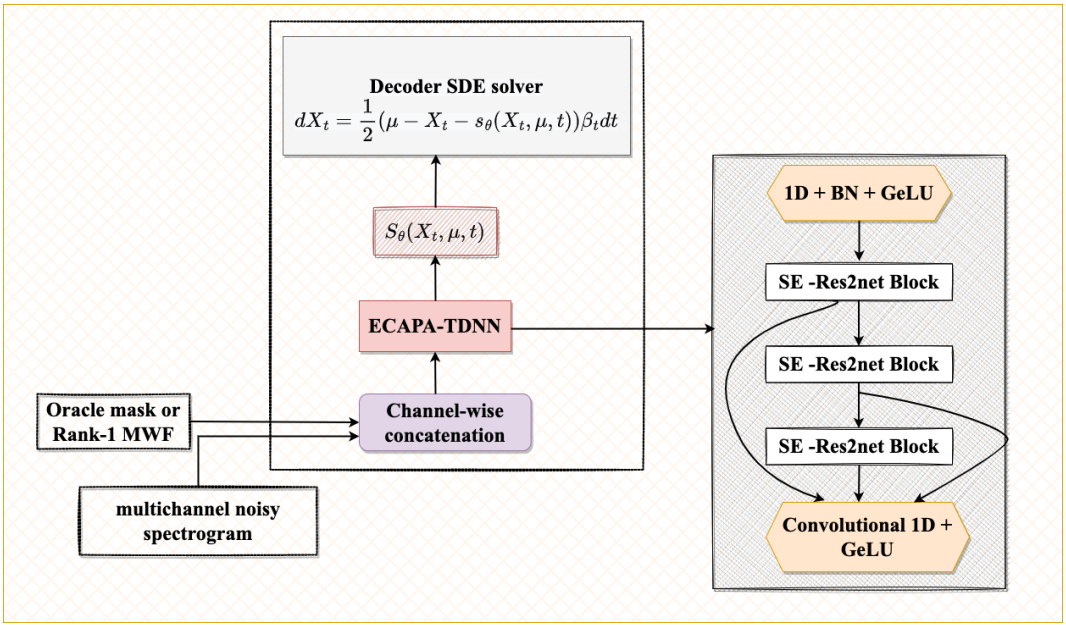

Figure 5.3: Graphical representation of Diff-Estimator. The Oracle mask or Rank-1 MWF and multi-channel noisy spectrogram first go to channel-wise concatenation to the diffusion network implemented using the encoder network of ECAPA-TDNN. The diffusion network learns to estimate the gradient of log probability density from each diffusion step. Later, solving SDE provides the desired data distribution as output. The main role of the ECAPA-TDNN network is to realize the diffusion decoder network, which has the sole purpose of estimating the gradient of log probability density.

mean of noise distribution is defined by the conditioning network's output. Additionally, we used MSE loss to ensure the training process's stability and provide smooth global optima in the optimization process.

We use scoring-based DPM, which uses Fisher divergence to define the diffusion loss [Song et al., 2020a]. We use Fisher divergence to minimize the divergence between the gradient of the log density of the noise and the gradient predicted by the DPM-based U-net decoder s_θ. Thus, diffusion loss enables s_θ to generate a better estimate of reverse trajectories of the forward diffusion process. The diffusion loss can be explained using the equation 2.40.

Inference

In the inference phase, enhancement is performed by the reverse diffusion process. We provide the noisy multi-channel Mel spectrogram as input to the conditioning network, which gives the output μ. After obtaining μ from the conditioning network, it is parameterized with latent variable sampled from Gaussian distribution. The mean of the Gaussian distribution is defined through the conditioning network's output μ, thus creating conditional Gaussian distribution depending on the output of the conditioning network μ. Afterwards, for given parameterized μ as input to the DPM-based decoder, the decoder performs a reverse diffusion process to transform latent variables into estimates of the target Mel spectrogram. Thus noise distribution is parameterized through the conditioning network's output μ. The reverse diffusion process reconstructs the Mel spectrogram of clean speech by gradually removing the noise sampled from the distribution $\mathcal{N}(\mu, I)$.

5.2.2 Diff-Estimator

We proposed to use DPMs for estimating either time-frequency masks or Rank-1 MWF and named our approach Diff-Estimator as shown in Figure 5.3. We developed two variants of the Diff-Estimator based on the type of output the network produces. Firstly, Diff-Estimator learns to estimate the Oracle mask of the clean speech given a multi-channel noisy spectrogram as input. While in the second case, Diff-Estimator is designed to compute the Rank-1 MWF given a multi-channel noisy spectrogram as input. The main objective of the Diff-Estimator is to learn the change in distribution from data to noise required to estimate the time-frequency mask or Rank-1 MWF. Like GradSE, Diff-Estimator incorporates a scoring function to realize the diffusion process. The Diff-Estimator system conducts speech enhancement on the amplitude of the spectrogram of noisy multi-channel speech.

The Diff-Estimator consists of only the diffusion decoder and does not use the conditioning network. This is because the purpose of the conditioning network is to make the terminal noise distribution closer to the desired/target data. However, in the case of the Diff-Estimator, the goal is not to generate a sample that is closer to the target distribution but rather to estimate a mask or compute Rank-1 MWF based on the input spectrogram and does not need to generate a complete sample. Therefore, the Diff-Estimator is designed as a conditional generative model.

We made two modifications to the DPM-based decoder compared to the DPM-based decoder of GradSE, as can be seen in Figure 5.3. The purpose of the modifications is to improve the accuracy and robustness of the DPM-based decoder for processing multi-channel noisy signals. The first modification involves replacing the U-net-based diffusion decoder with the encoder architecture from the ECAPA-TDNN system. This change may improve the ability of the decoder to extract relevant features from the input spectrogram and generate more accurate masks for speech enhancement. The second modification involves incorporating 1-D Squeeze-Excitation ResNet blocks and multi-layer output aggregation and summation. These changes allow the network to better capture the temporal context of the input spectrogram and adjust the channel weights based on the global features of the signal. This can lead to improved performance in separating speech from noise in multi-channel recordings.

The Diff-Estimator conducts a forward diffusion process by iteratively deconstructing the Oracle mask into a terminal noise distribution, which is parameterized using a noisy multi-channel spectrogram. In the training phase, the Diff-Estimator minimizes the scoring function in the forward diffusion process and learns to deconstruct the mask of clean speech to noise distribution. For training the Diff-Estimator, we used a single loss function which defines the scoring function as a diffusion loss term. In the reverse diffusion process, the Diff-Estimator initially samples noise from a distribution that is parameterized through a noisy multi-channel spectrogram. The ECAPA-TDNN-based diffusion model then uses a scoring-based function to predict the reverse trajectories of the forward diffusion process.

At the inference time, a noisy multi-channel spectrogram is given as input to ECAPA-TDNN, which then conducts T reverse diffusion steps to match the trajectories of forward diffusion in reverse order, thus enabling to compute the mask or Rank-1 MWF for a given multi-channel noisy spectrogram.

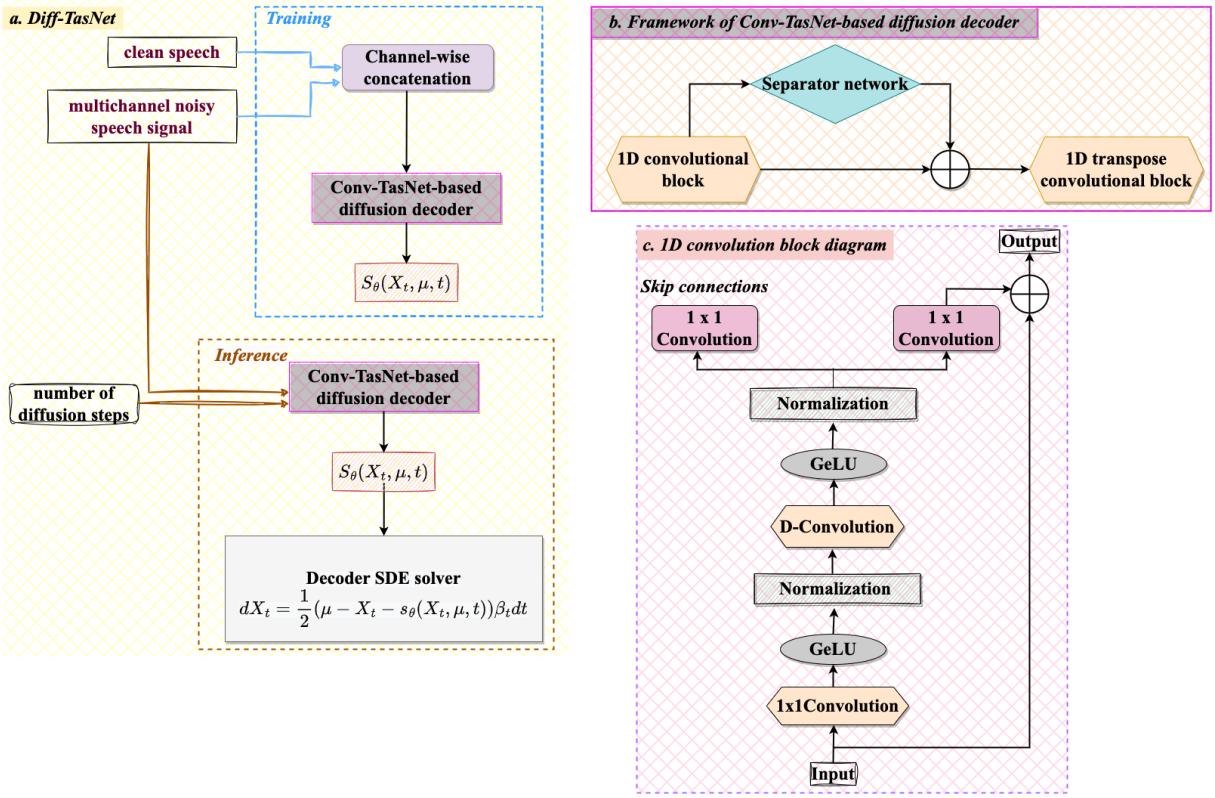

Figure 5.4: Architecture of Diff-TasNet. **(a)** We give the clean speech and multi-channel noisy speech signal as input to the Diff-TasNet during the training phase. **(b)** Illustration of the framework of convolution-TasNet-based diffusion decoder, which is consist of a 1D convolutional block, 1D transpose convolutional block and a separator network. **(c)** Illustration of the design of 1-D convolutional block used as a basic module in implementing conditioning network and diffusion decoder of Diff-TasNet.

5.2.3 Diff-TasNet

This section explains Diff-TasNet, a DPM-based time-domain speech enhancement technique for speaker verification. Diff-TasNet consists of a Conv-TasNet-based decoder. The core idea behind Diff-TasNet is to replace the U-net-based diffusion model with Conv-TasNet-based neural network architecture, as shown in Figure 5.4. The Conv-TasNet consists of an encoder, a separation module and a decoder. The Conv-TasNet model utilizes a linear encoder network to create an encoder representation optimized for speech enhancement or source separation tasks. After that, the masks are estimated using a temporal convolutional network (TCN) consisting of stacked 1-D dilated convolutional blocks, which allows the network to model the long-term dependencies of the speech signal while maintaining a small model size. However, unlike Conv-TasNet, Diff-TasNet excludes the decoder network and only estimates the masks in the time domain, which is interpreted as a scoring function to be learned during the training phase. We used the encoder and separation modules of multi-channel Conv-TasNet to estimate the scoring function to conduct forward and reverse diffusion processes. Diff-TasNet allows time-domain multi-channel speech enhancement, unlike U-net-based speech enhancement systems. Additionally, we replaced the ReLU activation functions in Conv-TasNet with GeLU (Gaussian error linear unit) to ensure the training stability and make output distribution throughout the layers

centred towards the Gaussian distribution. In contrast to previously proposed diffusion models for speech enhancement, Diff-TasNet inherently considers the phase information in the time-frequency domain.

Contrary to GradSE and Diff-Estimator approaches that rely on the spectrogram representation, Diff-TasNet directly performs speech enhancement using the raw time-domain multi-channel signal as input. Thus, to ensure the stability of the training process and to have a warm start for network parameters, we followed a two-stage training approach. First, we pre-trained the Diff-TasNet model as a generative autoencoder network, where a clean speech signal is iteratively deconstructed, and terminal noise distribution is conditioned using the same clean speech signal. In the second part of training, the multi-channel noisy speech signal is provided along with a clean speech signal as additional channel information to Diff-TasNet. This additional channel information is used to deconstruct it into a noise distribution in the forward diffusion process. Noise is sampled from the Gaussian distribution parameterized by the multi-channel noisy speech signal during the forward diffusion process.

To conduct the multi-channel speech enhancement, we provide a multi-channel noisy speech signal as an input to a Diff-TasNet network, where the diffusion model generates the reverse trajectories of the forward diffusion process. These reverse trajectories of the forward diffusion process are computed to transform the noisy multi-channel signal to an estimated clean speech signal by solving the SDE describing the dynamics of the reverse diffusion with a first-order Euler-Maruyama scheme.

5.2.4 Diff-Filter

In Chapter 4, it is illustrated that Rank-1 MWF improves speaker verification tasks compared to using other filtering techniques, such as GEV, or beamforming techniques, such as MVDR. This section presents a novel way to train a multi-channel speech enhancement system as a diffusion model-based filtering method named Diff-Filter. Diff-Filter is a scoring-based DPM where Conv-TasNet architecture is utilized for conducting the diffusion process. The proposed Diff-Filter model is an extension of Diff-TasNet and comprises a diffusion network and conditioning network as shown in Figure 5.5.

The Rank-1 assumption in MWF involves estimating the covariance matrices of speech and noise signals using only the sample mean of the input signals. This allows MWF to compute a filter that separates speech and noise components based on eigenvectors or generalized eigenvectors [Serizel et al., 2013]. Rank-1 MWF provides direct control over the level of speech distortion and noise reduction performance by adjusting the filter coefficients. This allows for a flexible trade-off between speech distortion and noise reduction performance. In addition to that, the Diff-Filter model uses a conditioning network to estimate the speech and noise signals given the noisy multi-channel signals required for the Rank-1 MWF filter to generate a clean speech signal. The estimated speech and noise signals are then concatenated with the noisy multi-channel signal and the Rank-1 MWF clean speech signal on a channel-wise basis as input to the diffusion decoder network to generate the filtered clean speech signal. We used the estimates of clean and noise signals as additional input to the diffusion decoder for conditioning the sampling process from terminal distribution aware of noise to be removed from the noisy multi-channel signal.

In the training phase, the forward diffusion process is conducted by iteratively deconstructing Rank-1 MWF clean speech signal to the terminal distribution defined by noisy

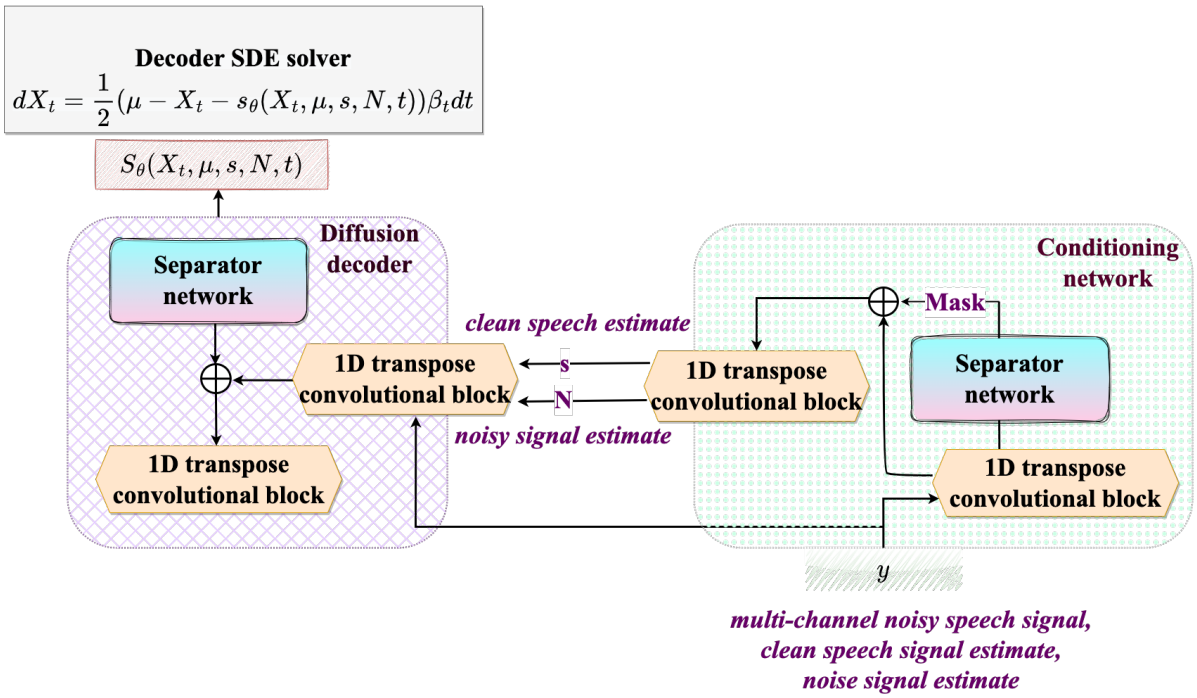

Figure 5.5: Architecture of Diff-Filter consists of a diffusion network and a conditioning network. We provide three inputs: multi-channel noisy speech signal denoted by y, clean speech signal estimate denoted by $\check{s}$ and noise signal estimate denoted by $\check{n}$ as a time-domain signal representation to output the Rank-1 MWF filter estimate for a clean speech signal. The clean speech signal and noise signal are channel-wise concentrated with noisy multi-channel signal along with Rank-1 MWF clean speech signal in the diffusion network.

multi-channel signal. Furthermore, terminal distribution is also conditioned with estimates of the clean speech signal and noise signal provided by the conditioning network. The usage of clean speech estimate and noise estimate in the diffusion process assists in conducting noise-aware speech enhancement. In the forward and reverse diffusion process, terminal noise distribution is defined as $\mathcal{N}(\mu, I)$, where the mean is μ, and I is unit variance. We parameterized the mean μ of terminal noise distribution of the diffusion process using noisy multichannel input, y.

We trained the Diff-Filter model using a two-stage training process. First, we conditioned the diffusion encoder with a target clean speech signal, a target noise signal, and a noisy multi-channel signal. We used pre-trained Conv-TasNet as an external conditioning network to compute the estimates of the clean speech signal, s, and noise in time-domain representation, N. The main purpose of pre-training the diffusion network is to ensure that diffusion model parameters converge in optimal minima direction using target clean speech and noise. In the second stage of training, we used the clean speech and noise estimated by the Conv-TasNet-based conditioning network. The Diff-Filter is trained using two loss functions, diffusion loss and SI-SDR loss. The diffusion loss is defined by Fisher divergence as a way to compute the scoring function. The second loss function, SI-SDR loss, is applied to the output of the conditioning network to ensure that the diffusion model ingrains the intrinsic information about clean speech estimate and noise estimate in time-domain representation. During the second-stage training, the model parameters of the conditioning network are also updated, which allows for improvement in the quality

of the estimated clean speech signal and estimated noise signal.

In the inference phase, multi-channel noisy signals and estimates of speech and noise signals obtained from the conditioning network are provided to the diffusion decoder to estimate the reverse trajectories of forward diffusion. The reverse diffusion process iteratively reconstructs the Rank-1 MWF filter output of clean speech by sampling latent variables from conditional terminal distribution.

5.2.5 Experimental setup

This section provides the experimental setup of the DPM-based speech enhancement models. The acoustic set-up remains the same as before. In all the diffusion-based models, Fisher divergence is incorporated to provide the scoring function, which minimizes the divergence between the gradient of the log density of the noise and the gradient predicted by the ECAPA-TDNN-based diffusion model. This diffusion loss enables the ECAPA-TDNN model to learn better estimates of reverse trajectories of the forward diffusion process. We conduct 20 reverse diffusion steps for all the models during inference.

GradSE

We used a convolutional recurrent neural network (CRNN) to implement the encoder network. CRNN network comprises a convolutional layer, batch normalization, ReLu activation function, and LSTM layers. The DPM-based decoder network is the U-net network from Ronneberger et al., 2015. We extract 40 dimensional Mel spectrogram features using the torchaudio library with a window length of 400 samples, hop size of 160, and 512 FFT length. For GradSE, the CRNN encoder is implemented using a 2D convolutional block of kernel size 3×3, a stride of 1, and padding of 1 with 3 input channels and a single output channel. We used 4 LSTM layers of 40 hidden dimensions. The encoder output is concatenated channel-wise and provided to the DPM-based decoder. GradSE is trained for 500 epochs, using a batch size of 32 and a learning rate of $1e^{-4}$.

Diff-Estimator

We compute the STFT for speech signals with an FFT length of 512 samples, a hop length of 256 samples and a Hanning window. During the training, we used the ideal ratio mask to compute the clean speech mask from the target clean speech and target noise signals. We trained the Diff-Estimator system for 1000 epochs using a batch size of 32 and a learning rate of $1e^{-4}$. We used the Adam optimizer to train the diffusion model based on the ECAPA-TDNN network. For implementing the ECAPA-TDNN network, we used 512 channels in all the convolutional layers. The dimension of the bottleneck in SE–Res2net Block and the attention module is set to 128. The scale dimension in SE–Res2net Block is set to 8.

Diff-TasNet

The diff-TasNet is trained using diffusion loss defined by Fisher divergence to compute the scoring function. In training, we provided speech segments of a fixed length of 4 seconds of the speech signal. Therefore, diff-TasNet is trained as a denoising autoencoder. We trained the Diff-TasNet for 100 epochs with a learning rate of 1e-2 with a weight decay of 1e-4 after every 5 epoch. We used Adam optimizer with a batch size of 4. For the

implementation of the diffusion decoder network, we used the Conv-TasNet framework and replaced Parametric Rectified Linear Unit (PReLU) activation function with the Gaussian Error Linear Unit (GeLU). The Conv-TasNet consists of 512 filters in the convolutional block and transpose convolutional block, 20 lengths of filters, 256 channels in a bottleneck, and the residual paths 1×1 convolution blocks. Each convolutional block's kernel size is set to 3, and the number of convolutional blocks in each repeat is 8. We adopted global layer normalization (GLN) with a non-causal strategy for Diff-TasNet implementation.

Diff-Filter

We used two loss functions, namely SI-SDR loss and fisher divergence loss. The SI-SDR loss is used on the output of the conditioning network to estimate the clean speech and noise to be given as additional input to the diffusion network. We set the weight of 0.001 on SI-SDR loss. Then, we increased the initial weight additively by 0.0001. For the first stage of the training, we first trained the network for 100 epochs with a learning rate of 1e-2 and reduced the learning rate over the epochs with a factor of 0.85 after every 5 epoch. We used Adam optimizer with a batch size of 2. In the second stage of training, the Diff-Filter is trained with a learning rate of $1e-4$ for 500 epochs.

We used Conv-TasNet architecture to develop the diffusion network and the conditioning network. We replaced the Parametric Rectified Linear Unit (PReLU) activation function with the Gaussian Error Linear Unit (GeLU). The implementation of networks using Conv-TasNet (both diffusion decoder and conditioning network) includes 512 filters in the convolutional block and transpose convolutional block, 20 lengths of filters, 256 channels in a bottleneck, and the residual paths 1×1 conv blocks. Each convolutional block's kernel size is set to 3, and the number of convolutional blocks in each repeat is 8. We adopted global layer normalization with a non-causal strategy for Diff-TasNet implementation. To ensure a stable learning process, we used gradient clipping with a maximum L2-norm of 5.

5.2.6 Speaker verification

We use the ECAPA-TDNN model architecture introduced by Desplanques et al., 2020. Besides squeeze and excitation block, the attention module of ECAPA-TDNN is set to 128. The scale dimension in Res2Block is set to 8. We extracted 256 dimension speaker embedding with the ECAPA-TDNN network. We trained the ECAPA-TDNN network on VoxCeleb1 [Nagrani et al., 2017] and VoxCeleb2 [Chung et al., 2018] datasets with a cyclic learning rate varying between $1e-8$ and $1e-3$ using the triangular policy with Adam optimizer. The ECAPA-TDNN network is trained with angular margin softmax with a margin of 0.3 and softmax pre-scaling of 30, 100k iterations. Mel spectrogram features of 40 dimensions as input to the ECAPA-TDNN network were extracted using the same procedure as for GradSE. We used a cosine scoring system for verification purposes applied to the extracted embeddings.

5.3 Results

We give the multi-channel signal as input to the diffusion-based speech enhancement systems and output the single-channel signal, which is further given as input to the speaker verification system. The performance of the DPM-based speech enhancement techniques,

both in terms of speech enhancement and as pre-processing to the speaker verification, is measured using three objective measures in terms of SIR (Signal-to-Interference Ratio) and SDR (Signal-to-Distortion Ratio) and EER (Equal error rate). For comparison, we also consider comparing the performance of the proposed techniques with the speech enhancement techniques from Chapter 4. All metrics are presented with a 95 % confidence interval using the bootstrap algorithm. We consider different conditions corresponding to different steps in the acoustic propagation process: dry clean speech, reverberated clean speech, and noisy (mixture of reverberated noise and speech). We compute EER on these conditions, and the signals are estimated with different speech enhancement algorithms.

5.3.1 Analysis of preliminary experimentation

Table 5.1 shows the results of preliminary experimentation conducted on the RoboVoices eval1 dataset, which tested different diffusion-based model architectures. We use the conditioning network of Diff-Filter described in Section 5.2.4 with Diff-TasNet described in Section 5.2.3. The table lists different models that were used: Diff-Filter described in 5.2.4 without the conditioning network, GradSE explained in Section5.2.1, for estimating masks, the conditioning network of GradSE described in Section 5.2.1 with the Diff-Estimator 5.2.2. The goal was to assess the impact of architectural changes on the performance of the speaker verification system.

Table 5.1: % EER on RoboVoices eval1 dataset with change in the model architecture for the DPM-based speech enhancement approaches for SNRs 5 dB, 10 dB, and 20 dB. SE refers to speech enhancement, and SV to speaker verification.

SE	SV	5	10	20
Diff-TasNet with conditioning network of Diff-Filter	ECAPA-TDNN	10.1	8.8	7.5
Diff-Filter without conditioning network	ECAPA-TDNN	10.0	8.1	7.0
GradSE as mask estimator	ECAPA-TDNN	10.6	9.1	8.0
Diff-Estimator with conditioning network of GradSE	ECAPA-TDNN	10.5	9.5	8.0

The results show that the Diff-TasNet model without the conditioning network performed better than using the conditioning network of Diff-Filter (refer to Table 5.2). Similarly, Diff-Filter degrades the performance without using the conditioning network (refer to Table 5.2). We also experimented with GradSE for estimating the time-frequency masks. We can see the performance degradation by GradSE Compared to Diff-Estimator. Diff-Estimator using the conditioning network of GradSE doesn't provide any significant performance as well.

5.3.2 Impact of diffusion-based pre-processing

Table 5.2 presents the results of different diffusion-based pre-processing approaches on the RoboVoices eval1 dataset. We reported the results for 3 different SNR (signal-to-noise ratio) conditions. We trained the models on the training set of the RoboVoices dataset (refer to Chapter 3). The first three rows of the table show the performance for dry clean speech, reverberated clean speech, and unprocessed noisy speech. These are used as reference signals to compare the performance of the speech enhancement approaches.

Table 5.2: SIR (dB), SDR (dB), and % EER on the simulated RoboVoices eval1 dataset concerning various SNR (dB). As it is not possible to compute SIR and SDR on Mel spectrograms, these metrics are not available for GradSE, which estimates Mel spectrograms. The average confidence interval is 0.1.

	5			**10**			**20**		
Pre-proces./SNR	**SIR**	**SDR**	**EER**	**SIR**	**SDR**	**EER**	**SIR**	**SDR**	**EER**
Dry clean speech	—	—	5.6	—	—	5.6	—	—	5.6
Reverb. clean speech	31.5	6.7	7.5	31.5	6.7	7.5	31.5	6.7	7.5
Unprocessed (Noisy)	12.9	0.8	11.2	15.1	2.0	9.2	17.4	5.0	7.8
GradSE	—	—	10.2	—	—	8.5	—	—	7.2
Diff-Estimator	21.1	5.2	10.4	20.8	5.9	8.9	26.3	6.2	7.5
ConvTasnet	21.6	5.3	10.2	23.2	6.1	8.6	27.5	6.6	7.1
Diff-TasNet	22.0	5.4	9.9	23.5	6.4	8.4	27.6	6.7	7.0
Diff-Filter	**22.4**	**5.6**	**9.7**	**23.7**	**6.6**	**7.8**	**27.7**	**6.8**	**6.5**

The remaining rows of the table show the performance of the proposed diffusion-based speech enhancement approaches.

The SIR and SDR values for unprocessed noisy speech are very low, indicating that there is significant interference and distortion in the signal. The EER is high for SNR 5 dB and 10 dB for this signal, indicating that the quality of the signal is low. But not for the 20 dB SNR condition. The proposed diffusion-based speech enhancement approaches improved the performance compared to unprocessed noisy speech, as evidenced by the higher SIR, SDR values and lower EER values. It is to be noted that the SIR and SDR for GradSE are not reported as classical speech enhancement evaluation metrics SIR, and SDR is computed on enhanced speech signal and original speech signal. But traditionally, input to speaker verification is either Mel filter bank features or Mel spectrogram, which is also the working domain of GradSE. Since the Mel spectrogram is a type of feature representation, it does not directly provide access to the individual signal, distortion, or noise components.

Diff-Filter achieved the best performance in terms of SIR, SDR, and EER across the SNR conditions indicating that it is the most effective at reducing interference and distortion in the signal. The EER values for this method are significantly lower than the other methods, indicating that it produces the highest quality enhanced speech compared to the other approaches. At the SNR of 5 dB, the average EER of 11.2% is obtained on the unprocessed signal, which is decreased to 9.7% using Diff-Filter. We can observe the same trend as the SNR increases. The speech enhancement and speaker verification performance are improved significantly by the diffusion-based models in all the SNR conditions.

5.3.3 Performance on SV-dedicated dataset

Table 5.3 presents the results on the MultiSV dataset. The MultiSV dataset provides evaluation trials with multiple conditions based on the properties of the enrollment segments to support various use cases [Mošner et al., 2022]. The trials contain multi-channel test recordings. Among four conditions, we experimented with two conditions that best represent our case scenario. We evaluated the MRE or multi-channel re-transmitted en-

rollment and MRE-hard, which has hard background noise. We trained all the models on the training set of the MultiSV dataset. The results are averaged and presented in terms of EER. The third section in the table shows the performance of the speaker verification model when the speech signal is enhanced using the time-frequency masks estimated using diffusion-based approaches. And the fourth section in the table shows the performance of the speaker verification model when the speech signal is enhanced using multi-channel filters computed with diffusion-based approaches.

Table 5.3: % EER on the MultSV Eval dataset. Mask and filter in the table refer to the speech enhancement output, which is further given as input to the speaker verification. The output of GradSE is the Mel spectrogram. The average confidence interval is 0.1.

	SE	SV	MRE	MRE-Hard
	Unprocessed	ECAPA-TDNN	5.84	10.27
	Oracle Rank-1 MWF	ECAPA-TDNN	1.64	3.12
	GradSE	ECAPA-TDNN	3.77	4.52
Mask	Diff-Estimator	ECAPA-TDNN	3.63	4.57
Mask	Diff-TasNet	ECAPA-TDNN	**3.58**	**4.51**
Mask	Diff-Filter	ECAPA-TDNN	3.65	4.56
Filter	Diff-Estimator	ECAPA-TDNN	3.61	4.55
Filter	Diff-TasNet	ECAPA-TDNN	3.57	4.51
Filter	Diff-Filter	ECAPA-TDNN	**3.53**	**4.36**

We used the Oracle Rank-1 MWF from [Delebecque et al., 2022] as a reference signal to compare the speaker verification performance, which achieved an EER of 1.64% on MRE, and 3.12% on the MRE-hard. Table 5.3 shows that the signal quality degrades mostly in the MRE-hard condition with an EER of 10.27% without any pre-processing compared to the MRE condition. The high EER value in the unprocessed input for MRE-hard indicates the presence of high levels of noise and reverberation.

The diffusion-based models consistently improved the speaker verification performance in MRE and MRE-hard conditions. The performance of mask-based speech enhancement approaches (the three rows under Mask) is similar. With an EER of 3.58% in the MRE condition and 4.51%, Diff-TasNet obtained the best performance compared to the other approaches for the speech enhancement output mask.

The last three rows of the table show the results for the speech enhancement approaches with a filter output. With an EER of 3.53% in the MRE condition and 4.36% in the MRE-hard condition, Diff-Filter achieved the best performance for the speech enhancement output filter. The Diff-Filter improved the speaker verification performance by almost 6% over the unprocessed signal giving the speech enhancement output multi-channel filter as input to the speaker verification.

5.3.4 Validation on public dataset

Table 5.4 presented the results on different noise conditions of the VOiCES dataset. We evaluated our systems' performances on the publicly available VOiCES Eval challenge dataset [Richey et al., 2018] without re-training the system. The first section of the Table shows the EER on different noise conditions with Oracle Rank-1 MWF and the

unprocessed signal. The second section of the Table shows the EER of the proposed diffusion-based models.

Table 5.4: % EER on different noise conditions of the VOiCES Eval dataset. The average confidence interval is 0.2.

	Noise conditions			
Testing environment	**Clean**	**Babble**	**TV**	**Music**
Unprocessed	4.1	8.8	7.8	7.9
Oracle Rank-1 MWF	2.8	3.9	4.1	4.2
GradSE	3.9	6.7	6.2	6.6
Diff-Estimator	3.9	6.6	6.6	6.8
Diff-TasNet	3.9	6.3	6.1	6.2
Diff-Filter	**3.8**	**6.2**	**6.0**	**6.0**

We observe that diffusion-based models improve the speaker verification performance for all noise conditions, the largest gain being observed for babble noise, especially Diff-Filter, with an EER of 6.2%. Diff-Filter has achieved the best performance across all the noise conditions. The results show that while speech enhancement is effective in separating speech from noise, their performance is influenced by the type of noise. Although the proposed DPM-based systems are not trained on the VOiCES dataset, they have improved the EER across all the noise conditions, demonstrating our approach's efficacy even though the model was trained on a synthetic dataset generated for generic, possibly mismatched, and spatial scenarios.

5.4 Conclusion

This chapter presented various speech enhancement techniques based on scoring-based diffusion probabilistic models for speaker verification in noisy and reverberant environments. To improve the noise robustness of a far-field speaker verification system, we proposed using diffusion probabilistic models for multi-channel speech enhancement as a front-end processing. We opted for scoring-based diffusion models, which we implemented using SDE. The score-based diffusion models generate high-quality samples by mapping data to noise by gradually perturbing the input data in the forward process and learning to revert this process.

We proposed diffusion-based speech enhancement techniques to facilitate dealing with challenging acoustic environments, which severely distort the speech signal in far-field speaker verification. Our experimentation with diffusion-based models suggests that it can achieve high-quality enhancement using score-based DPM without degrading the quality and intelligibility of the signal, even in challenging scenarios.

We replaced the U-net-based decoder with the ECAPA-TDNN for the Diff-Estimator, incorporating a scoring function to compute the time-frequency masks. The channel-dependent frame attention of the statistics pooling layer and the additional skip connections of ECAPA-TDNN enables the network to concentrate on the important frames. The frame and statistics pooling layers integrate channel attention that uses a global context to enhance the performance even more. This can be observed in the results produced with the Diff-Estimator.

The performance of the Diff-Filter for computing multi-channel filtered clean speech shows that conditioning the estimates of the clean speech and the noise with the multi-channel noisy signal during the diffusion process assists in generating quality samples, further preventing signal degradation by maintaining intelligibility in the back-end speaker verification system.

We evaluated the performance of the speech enhancement and speaker verification systems using SDR, SIR, and EER metrics. The experiment showed that diffusion-based techniques improved speaker verification performance compared to the current state-of-the-art methods described in Section 4.4.2 across all datasets, even in challenging scenarios with different background noise and reverberation levels. The results showed that the Diff-Filter approach was the most effective and produced the best results while maintaining high speech quality.

Chapter 6

Joint optimization of speech enhancement and speaker verification

Contents

6.1 Introduction

In Chapter 2, we have discussed different noise interference affecting far-field speaker verification performance. It was pointed out that background noise can contain fluctuating or steady noise that significantly affects the speech signal quality. Different kinds of noise can corrupt the features of the speech signal, which can affect how well speaker verification algorithms function. Additionally, when the level of the background noise rise, the effectiveness of acoustic source separation techniques and speaker verification systems rapidly degrades. Section 2.5 and Chapters 4, and 5 discussed speech enhancement as one of the potential solutions to mitigate the distorted signal and improve the quality and intelligibility by removing all kinds of background noises and room reverberation.

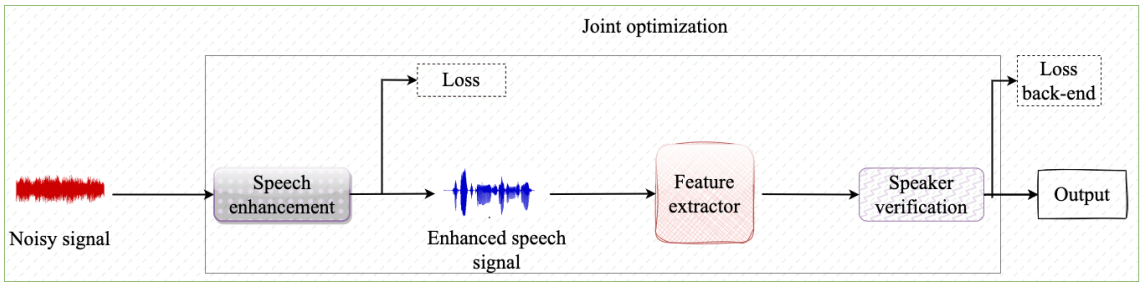

Figure 6.1: Schematic illustration of a generic jointly optimized speech enhancement and speaker verification system.

The speech enhancement model can be implemented individually and utilized as a pre-processing module to a speaker verification system to filter out the noisy and reverberated interference by generating a time-frequency mask or directly predicting the clean signal from the noisy input. However, the independent training of the speech enhancement model can result in mismatch problems. The primary reason for the mismatch issue is caused because the training target of the speech enhancement model is to reduce the noisy reverberated particles by generating a mask or a clean signal directly. The speech enhancement model not only filters out the noise interference from the noisy input but also corrupts some of the features required by the back-end speaker verification model. For instance, in the mapping-based speech enhancement model, learning the mapping from the noisy speech to the clean signal is the goal of the speech enhancement model. The loss function's reconstruction serves as the objective function's foundation. Because the speech enhancement model is trained independently, there is no constraint in the loss function to ensure that it will retain the features useful to the back-end model. As a result, speech enhancement systems frequently struggle to distinguish between a useful feature and noise interference in the enhanced signal. In contrast, there are no such restrictions on the features used for the back-end speaker verification model, for instance, speaker-related features for the back-end speaker verification system. According to Sadjadi and Hansen, 2010, Shon et al., 2019, Wang and Wang, 2016, the mismatch issue was caused by the speech enhancement module distorting some valuable features and effectively generating additional interference, adversely impacting the back-end model.

We jointly optimized the speech enhancement approaches proposed in chapters 4 and 5 and the speaker verification module to improve the performance as joint optimization helps in filtering out the noise and reverberation and retaining useful information. Few studies jointly optimized or integrated weighted prediction error (WPE) and some variants of beamforming using speaker embedding model for reducing the EER [Yang and Chang, , Taherian et al., 2019b, Mošner et al., 2018]. Yang and Chang, presents a multi-step approach for joint optimization. First, they train a DNN-supported MVDR beamformer and WPE on multi-channel speech signals. Then, they train an x-vector speaker embedding network on top of the enhanced speech features produced by the beamformer and WPE. Finally, they connect the beamformer and WPE to the x-vector and jointly train all three components. They build separate speaker embedding models using the output of the front end to compare their effectiveness for deep speaker modelling. Shon et al., 2019 tried to integrate speech enhancement and speaker verification modules into a single framework for speaker verification. The speech enhancement module filters out the corrupted features by generating a ratio mask and multiplying element-wise with the original spectrogram for speaker verification. However, they processed the speech en-

hancement module individually, and the speaker verification module was pre-trained and frozen during the training of the speech enhancement. This means that the two modules were not optimised jointly. Shi et al., 2020 used an attention mechanism, cascaded speech enhancement network and speaker recognition by jointly optimizing their parameters using a single loss function. The speech enhancement network uses dilated convolution and multi-stage attention blocks and takes the noise spectrogram as input. The speaker recognition network, known as SID-Net, uses residual convolution and multi-stage attention blocks and takes the enhanced spectrogram as input.

There are several benefits of jointly optimizing speech enhancement and speaker verification models:

i. Improved performance: By considering the interdependencies between speech enhancement and speaker verification, a joint optimization framework can achieve better performance than optimizing each task separately.

ii. Increased robustness: By considering the presence of noise or other distortions during training, the joint optimization framework can make the speaker verification model more robust to these factors.

iii. Enhanced feature representation: By optimizing both tasks together, the feature representation used for speaker verification can be enhanced, leading to improved performance.

In this chapter, we proposed two approaches for joint optimization based on noise reduction techniques with and without knowledge distillation loss. The first joint-optimized approach implements FaSNet, a neural beamforming technique and an ECAPA-TDNN-based speaker verification system. And the second approach is based on DPM-based multi-channel speech enhancement approaches and the ECAPA-TDNN-based speaker verification system. We jointly optimized the DPM-based speech enhancement models described in chapter 5, namely: (i) GradSE, (ii) Diff-Estimator, (iii) Diff-TasNet, and (iv) Diff-Filter with speaker verification. Additionally, we proposed using knowledge distillation (KD) loss with FaSNet described in Chapter 4 and GradSE. We propose to use the similarity-preserving knowledge distillation loss because it guides the network to produce similar activation for enhanced signals to clean speech signals. The similarity-preserving knowledge distillation technique minimizes the distance between speaker embeddings obtained from the proposed system and that of clean speech signals. Experimental evaluation shows significant improvement in performance by DPM-based models over state-of-the-art models in alleviating noisy-reverberated speech without degrading the signal quality. We describe the joint-optimized approaches in the following sections and discuss the experimental results.

6.2 Joint Optimization of speech enhancement and SV

We re-used the proposed architectures on speech enhancement from previous chapters and ECAPA-TDNN-based speaker verification for joint optimization. We fine-tuned the proposed pipeline as a single differentiable unit, combining multi-channel speech enhancement and speaker verification. We extract the last hidden output of the ECAPA-TDNN network as speaker embedding, where the ECAPA-TDNN network is trained for a classification task. Then, we provide the generated speaker embedding to the classifier to derive

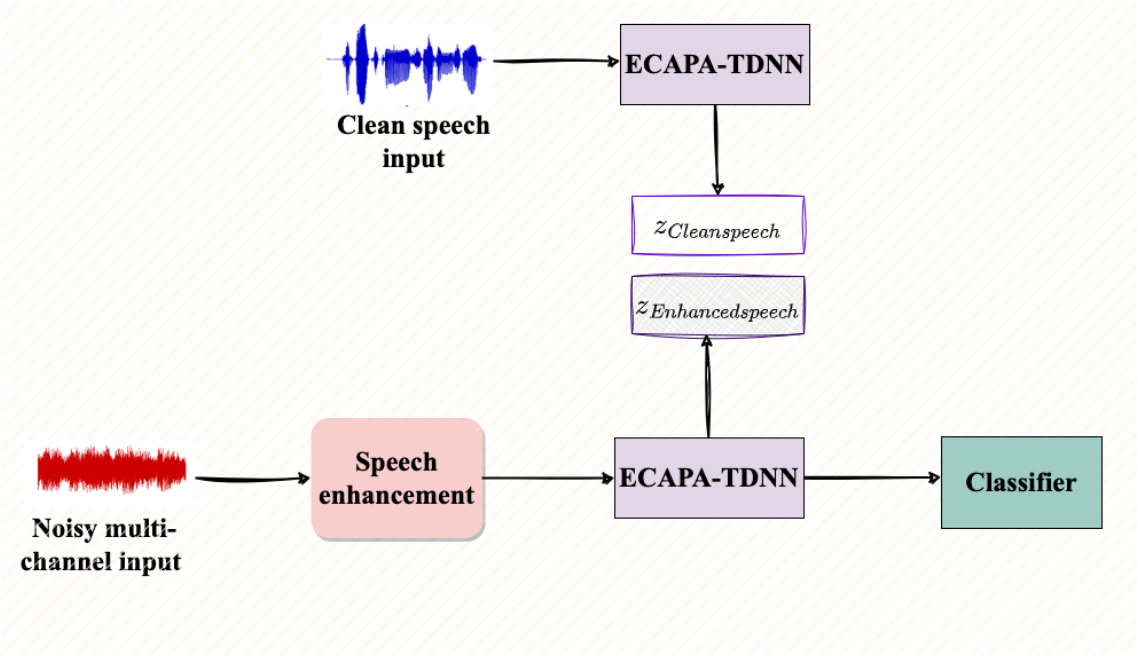

Figure 6.2: Joint optimization of speech enhancement and ECAPA-TDNN-based speaker verification. Embedding generated by ECAPA-TDNN for clean speech signal is considered the teacher network. Embedding generated by a joint optimized network on a multi-channel noisy signal is considered the student network.

the softmax probability distribution, which is later used for computing cross-entropy loss with target speaker labels. Therefore, for a given multi-channel noisy speech signal, the proposed architecture produces speaker embedding as a final output of the network. The fine-tuning procedure of joint optimization involves back-propagation of error gradients through speaker verification and speech enhancement networks in a single backpropagation pass. We used the jointly trained models for computing speaker embeddings for enrollment and test trials.

Now, we will detail the joint optimization for various speech enhancement systems. Figures 6.2 and figure 6.3 illustrate the architecture of the joint optimization of speech enhancement approaches and speaker verification with and without knowledge distillation loss.

6.2.1 FaSNet

We optimize the FaSNet-based speech enhancement model and the ECAPA-TDNN-based speaker verification system in a single framework. In section 4.2, we described the details of FaSNet for multi-channel speech enhancement. We have used scale-invariant signal-to-distortion ratio (SI-SDR) to train FaSNet and cross-entropy loss to train the speaker verification system. Cross-entropy computes the average difference between two probability distributions. For joint optimization, we have used the sum of both loss functions.

We give a noisy multi-channel signal to FaSNet, which emulates neural beamforming to provide output as a single-channel clean speech signal. For jointly optimizing with the ECAPA-TDNN-based speaker embedding network, we have given this enhanced speech signal to the torch-audio-based MFCC feature extractor. And subsequently, the ECAPA-TDNN network consumes this feature to learn the speaker representation as an embedding.

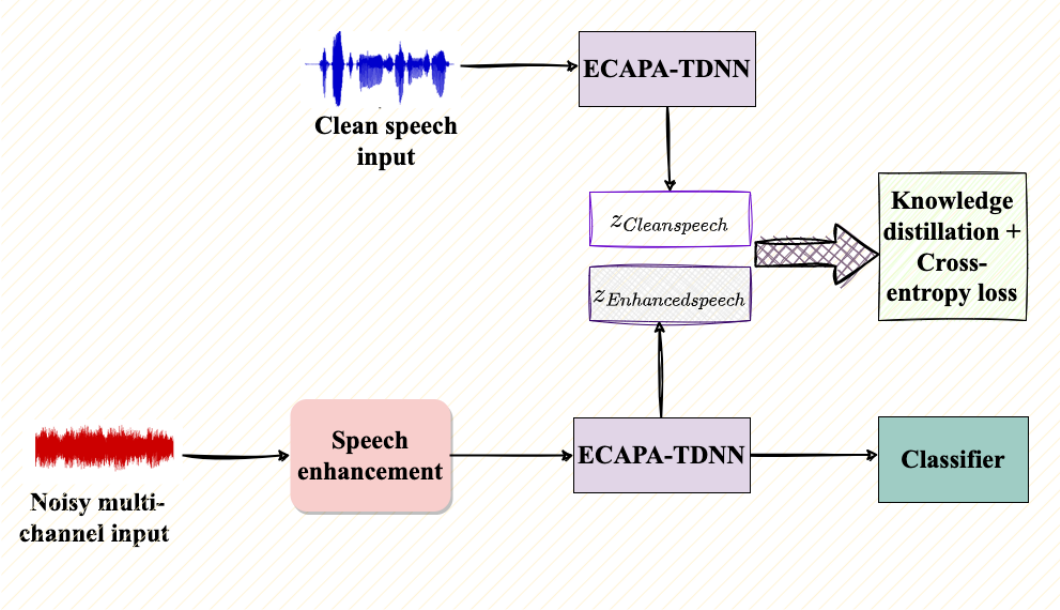

Figure 6.3: Joint optimization of multi-channel speech enhancement with speaker verification using knowledge distillation loss. Embedding generated by ECAPA-TDNN for clean speech Mel spectrogram is considered the teacher network. Embedding generated by a joint optimized network on a multi-channel noisy Mel spectrogram is considered the student network.

In the training phase of joint optimization, we used the sum of both the loss functions, weighted SI-SDR loss with a weight of 0.1 multiplied by SI-SDR loss and cross-entropy loss with a weight of 0.9. After conducting experiments with various weights, we found that 0.1 and 0.9 are the optimal weights on the loss function. We opted for SI-SDR loss to ensure training stability and guide the fine-tuning network to produce speech enhancement output. We use the Adam optimizer to train the jointly optimized network for 50 epochs with a batch size of 4 and a learning rate of 0.0001.

6.2.2 GradSE

As the DPM-based speech enhancement approaches were described in chapter 5. We will only describe the joint optimization process in the following sections.

We optimize the DPM-based multi-channel speech enhancement module GradSE with the ECAPA-TDNN speaker verification system in a single framework. We give the noisy multi-channel Mel spectrogram as input to GradSE. GradSE then performs a reverse diffusion process to remove the noise in input to reconstruct the target clean Mel spectrogram with a time horizon of 20 time-steps, which means the network conducts 20 reverse diffusion steps to reconstruct the target clean Mel spectrogram. The enhanced Mel spectrogram from GradSE is then passed through the ECAPA-TDNN network. Finally, the error gradients are passed through ECAPA-TDNN and GradSE in the back-propagation pass.

We jointly optimized both networks using cross-entropy loss on predicted labels by the classifier and target speaker labels and similarity preservation KD loss. We performed the joint optimization using the Adam optimizer and cyclic learning rate scheduler varying between 1e-3 and 1e-1 using the triangular2 policy. During the joint optimization process,

we trained the network with angular margin softmax with a margin of 0.4 and softmax pre-scaling of 30. We used a batch size of 64 and trained for 20k iterations. We opted for 20 steps for reverse diffusion for speech enhancement after analyzing the trade-off between performance on speaker verification and inference speed.

6.2.3 Diff-Estimator

Diff-Estimator is a multi-channel speech enhancement system based on the diffusion process for estimating masks for clean speech. We provided a spectrogram of noisy multi-channel speech as an input to Diff-Estimator. After conducting 20 reverse diffusion steps, the ECAPA-TDNN-based decoder network provides output as an estimated mask for clean speech. The estimated mask is applied to the noisy multi-channel speech to produce a clean speech estimate, which is then transformed into a single-channel clean speech signal. Thereafter, Mel spectrogram features are computed and given to the speaker verification system. As all the intermediate outputs are processed using the torchaudio library, this enables passing the error gradients through both SE and speaker verification systems. In the joint training phase, we used only a single loss function, cross-entropy loss applied on speaker labels predicted from the speaker verification system. For this fine-tuning process, both systems are jointly trained for 500 iterations with a batch size of 16 and a learning rate of 0.0001 and Adam optimizer.

6.2.4 Diff-TasNet

Diff-TasNet-based multi-channel speech enhancement utilizes Conv-TasNet as a decoder network for implementing a diffusion process to remove noise iteratively from a multi-channel noisy signal. We use Conv-TasNet as a diffusion decoder as it takes time domain signal as input. Diff-TasNet first performs reverse diffusion to provide an estimate of the clean speech signal with 20 reverse diffusion processes. Afterwards, Mel spectrogram features are extracted using torchaudio and provided as input to the ECAPA-TDNN-based speaker verification system for jointly training SE and speaker verification. We jointly trained both systems for 300 iterations with a batch size of 4 and a learning rate of 1e-4 with adam optimizer. We opted for a single loss function, cross-entropy loss, so that the speech enhancement system adjusts the network parameters focused on learning speaker representation under noisy conditions.

6.2.5 Diff-Filter

The Conv-TasNet network was used in Diff-Filter to create both diffusion and conditioning networks. More information about this can be found in Section 5.2.4. The implementation of the Conv-TasNet networks involved using 512 filters in both the convolutional and transpose convolutional blocks, with 20 lengths of filters, 256 channels in a bottleneck, and 1×1 convolutional blocks for residual paths. Each convolutional block used a kernel size of 3, and there were 8 convolutional blocks in each repeat. To ensure stable learning, global layer normalization was used with a non-causal strategy, and gradient clipping was used with a maximum L2-norm of 5.

For joint optimization with ECAPA-TDNN-based speaker verification, the pre-trained Diff-Filter conducted reverse diffusion to produce a filtered enhanced clean speech signal with 20 diffusion steps for a given input as a multi-channel noisy signal. Mel spectrogram

features were then extracted using torchaudio and provided as input to the ECAPA-TDNN-based speaker verification system for training both SE and speaker verification jointly. The systems were trained for 300 iterations with a batch size of 4 and a learning rate of 1e-4, using the adam optimizer. To allow the speech enhancement system to focus on learning speaker representation under noisy conditions, a single cross-entropy loss function was used.

6.3 Transfer learning with Knowledge distillation loss

Knowledge distillation is used as a model compression technique in machine learning, where a smaller model, called the student model, is trained to mimic the behaviour of a larger, more complex model, called the teacher model. The goal is to extract the knowledge of the teacher model and transfer the knowledge learned by the teacher model to the student model, allowing the student model to achieve similar performance with fewer parameters and less computational resources [Hinton et al., 2015]. To train the student model, a loss function is used to measure the difference between the output of the student model and the output of the teacher model. The most common loss function used in knowledge distillation is the knowledge distillation (KD) loss, also known as the soft target loss. Traditionally KD loss is used for improving inference speed and reducing the model parameters. Moreover, the distillation loss reduces the divergence between the output distribution of the teacher network and the student networks' output distribution.

The KD loss is defined as the mean squared error (MSE) between the soft targets produced by the teacher model and the softmax outputs produced by the student model. Soft targets are the probabilities produced by the teacher model for each class, which are typically smoother and more informative than the one-hot encoded labels used for training the original teacher model.

Formally, the knowledge distillation loss can be written as:

$$L = \alpha * T^2 * KL(y_{teacher}, y_{soft}) + (1 - \alpha) * MSE(y_{teacher}, y_{student}) \tag{6.1}$$

where $y_{teacher}$ is the soft targets produced by the teacher model, y_{soft} is the softmax outputs produced by the student model, KL is the Kullback-Leibler divergence, MSE is the mean squared error, T is a temperature parameter used to soften the probabilities produced by the teacher model, and α is a weighting factor that balances the importance of the two terms in the loss function.

By minimizing the knowledge distillation loss, the student model learns to replicate the behaviour of the teacher model, effectively transferring the knowledge learned by the teacher model to the student model.

We have used knowledge distillation (KD) loss with two jointly optimized systems: FaSNet and GradSE. KD loss derives the information to minimize the distance between speaker embedding from noisy and clean speech signals. Therefore, KD loss enables proposed joint optimization to generate embeddings closer to that generated by clean speech.

6.3.1 Similarity-preserving knowledge distillation loss

Similarity-preserving knowledge distillation is a variant of the knowledge distillation technique that aims to preserve the similarity between the feature representations of the

teacher model and the student model, in addition to replicating the output probabilities. The idea behind similarity-preserving knowledge distillation is that the feature representations learned by the teacher model are likely to be informative and useful for the task at hand, and by that encouraging the student model to learn similar feature representations, it can improve its overall performance. To achieve this, a new loss function is used, called the similarity-preserving KD loss [Tung and Mori, 2019][8].

Similarity-preserving KD loss is motivated by the idea that semantically similar inputs tend to obtain similar activation patterns in a trained neural network. Similarity-preserving KD loss aims to use the pairwise activation similarities within each input mini-batch to supervise the training of a student network with a trained teacher network. The similarity-preserving KD loss requires the student network to maintain the pairwise similarities in its representation space rather than replicating the teacher network's representation space.

Formally, the similarity-preserving knowledge distillation loss can be written as:

$$L = \alpha * T^2 * KL(y_{teacher}, y_{student}) + (1 - \alpha) * \beta * ||f_{teacher} - f_{student}||^2 \tag{6.2}$$

where $y_{teacher}$ and $y_{student}$ are the output probabilities produced by the teacher and student models, respectively, $f_{teacher}$ and $f_{student}$ are the feature representations produced by the teacher and student models, respectively, KL is the Kullback-Leibler divergence, $||\ ||$ denotes the $L2$ norm, T is a temperature parameter used to soften the probabilities produced by the teacher model, α and β are weighting factors that balance the importance of the two terms in the loss function. We used knowledge distillation loss with FaSNet and GradSE systems, where we conducted experimentation with different values of α and β, where $\beta = (1 - \alpha)$. From Table 6.1, we observed that for training jointly optimized systems with KD loss, 0.9 as the optimal weight for α.

By minimizing the similarity-preserving knowledge distillation loss, the student model learns to produce similar output probabilities and feature representations as the teacher model, achieving similar performance while using fewer resources.

We apply similarity-preserving KD loss to facilitate the jointly optimized system to generate embeddings closer to those generated by the clean speech by guiding the network to produce similar activation for noisy signals to that of clean speech signals. Thus, similarity-preserving KD loss derives the information to minimize the distance between speaker embeddings obtained from the proposed system ($z_{Enhanced-speech}$) and that of clean speech signals ($z_{Clean-speech}$), as shown in Figure 6.3. We used the same pre-trained ECAPA-TDNN model as a teacher network with embeddings obtained on clean speech signals and our proposed joint optimization of the GradSE and ECAPA-TDNN as the student network. The similarity-preserving KD loss assists the student model in matching the performance of enhanced signal embedding to the embeddings from a clean speech signal. Thus, the KD technique allows the proposed system to learn the robust latent space of speaker representation in noisy scenarios. In addition to GradSE, we used FaSNet based speech enhancement system for implementing a baseline system. We used the same model hyper-parameters as mentioned in the FaSNet system.

Table 6.1: Experimental performance evaluation of α weight parameter in knowledge distillation with GradSE and FasNet SE systems and ECAPA-TDNN-based speaker verification system.

Weight, α	**EER, FaSNet + KD Loss**	**EER, GradSE + KD Loss**
0.25	10.9	10.3
0.75	10.2	9.7
0.9	**9.9**	**9.2**
0.95	10.1	9.5

6.4 Results

To evaluate the proposed jointly optimized models' effectiveness and noise robustness. We conducted several experiments using both a simulated dataset and the real-recorded dataset. We computed the results on RoboVoices eval1 (refer to section 3.3.3), MultiSV (refer to section 3.2.1) and VOiCES (refer to section 3.2.1). We evaluate speech enhancement performance using the SIR and SDR metrics and speaker verification performance using EER. It is to be noted that the SIR and SDR for GradSE are not reported as classical speech enhancement evaluation metrics SIR, and SDR is computed on enhanced speech signal and original speech signal. But traditionally, input to speaker verification is either Mel filter bank features or Mel spectrogram. Therefore, we provided a noisy multi-channel Mel spectrogram to ease the joint optimization of speech enhancement and speaker verification systems. The Conv-TasNet model is implemented from [Mošner et al., 2022] and treated as a baseline. For comparative purposes, we also report the performance of non-joint optimized models from the previous chapters.

Table 6.2: SIR (dB), SDR (dB), and % EER on the simulated RoboVoices dataset concerning various SNR (dB). J. optim. in the table refers to joint optimization. The average confidence interval is 0.1.

		5			**10**			**20**		
	Pre-pro./SNR	**SIR**	**SDR**	**EER**	**SIR**	**SDR**	**EER**	**SIR**	**SDR**	**EER**
	Dry clean speech	—	—	5.6	—	—	5.6	—	—	5.6
	Reverb. clean speech	31.5	6.7	7.5	31.5	6.7	7.5	31.5	6.7	7.5
	Oracle Rank-1 MWF	22.7	6.0	5.9	24.7	7.2	5.9	28.2	7.5	5.9
	Unprocessed (noisy)	12.9	0.8	11.2	15.1	2.0	9.2	17.4	5.0	7.8
	Conv-TasNet	21.6	5.3	10.2	23.2	6.1	8.6	27.5	6.6	7.1
	Diff-Estimator	21.1	5.2	10.4	20.8	5.9	8.9	26.3	6.2	7.5
	Diff-TasNet	22.0	5.4	9.9	23.5	6.4	8.4	27.6	6.7	7.0
	Diff-Filter	22.4	5.6	9.7	23.7	6.6	7.8	27.7	6.8	6.5
J. optim.	Conv-TasNet	22.4	5.5	9.3	23.8	6.5	7.8	27.8	6.8	7.2
J. optim.	Diff-Estimator	22.3	5.4	9.6	21.9	6.2	7.9	26.7	6.4	7.1
J. optim.	Diff-TasNet	22.7	5.9	8.8	24.0	6.8	7.5	28.0	7.1	6.6
J. optim.	Diff-Filter	**23.0**	**6.0**	**8.1**	**24.1**	**6.9**	**7.5**	**28.1**	**7.3**	**6.4**

[8]https://github.com/AberHu/Knowledge-Distillation-Zoo

6.4.1 Impact of SNR

Table 6.2 shows the results on the RoboVoices eval1 dataset at different SNR (5, 10 and 20 dB) levels. The EER for dry clean speech and reverberant clean speech is 5.6, while the EER for unprocessed (noisy) speech is higher at 11.2 for an SNR of 5 dB. This indicates that noise has a significant impact on speech enhancement performance. At 5 dB SNR, the Oracle Rank-1 MWF showed 9.8 and 5.2 of absolute improvement to the unprocessed signal in terms of SIR and SDR and 5.3% in terms of EER. The Oracle Rank-1 MWF is used as a reference signal to compare the performance of speech enhancement approaches and speaker verification.

The second section of the table presents the results for different individual speech enhancement approaches, including Diff-Estimator, Diff-TasNet, Diff-Filter and Conv-TasNet. As we move from Diff-Estimator to Diff-Filter, the SIR and SDR metrics improve, and the EER decreases, indicating better performance. Diff-Filter shows the best performance among all approaches, with the highest SIR and SDR and the lowest EER, indicating that it is the most effective method for enhancing speech in noisy environments.

The third section of the table presents the results for joint optimization of the speech enhancement approaches with ECAPA-TDNN-based speaker verification. We can see that joint optimization of speech enhancement and speaker verification has improved the performance across the evaluation metrics in all the SNR conditions than the individual approaches. At the SNR of 5 dB, mean unprocessed scores average 11.2% in terms of EER, which is decreased to 8.1% using the joint optimization of the Diff-Filter model with speaker verification, respectively. This shows that the proposed joint-optimized model is robust to low SNR scenarios where diffuse noise and reverberation are actively present. We can observe the same trend as the SNR increases.

Overall, joint optimization has the potential to improve the robustness and effectiveness of both speech enhancement and speaker verification systems, leading to better performance in real-world scenarios where noise and reverberation are present.

6.4.2 Impact of knowledge distillation loss

Table 6.3 provides the results of two speech enhancement approaches, FaSNet and GradSE, on the RoboVoices eval1 dataset, in terms of SIR, SDR, and EER. The evaluation is conducted on various SNR levels, 5 dB, 10 dB, and 20 dB. The second section presents the results of the individual approaches. The last section presents the results of the joint optimization of the speech enhancement approaches and ECAPA-TDNN-based speaker verification with and without KD loss.

It is important to note that GradSE works with Mel spectrograms, and it is not possible to compute the SIR and SDR using the Mel spectrogram; we have not reported the results with these metrics. Therefore, the table only provides the EER metric for GradSE.

GradSE achieves the best results for the EER metric. Although FaSNet performs better for SIR and SDR metrics, it obtained the worst performance in terms of EER as expected, which we have already seen in the other experiments (refer to the experiments in chapters 4 and 5). The results show that joint optimization of FaSNet and GradSE with ECAPA-TDNN-based speaker verification provides significant improvements over the individual approaches in terms of EER. Specifically, the EER values are significantly lower than the individual approaches on all SNR levels. This indicates that the joint optimization approach is effective in enhancing the performance of speaker verification.

Table 6.3: SIR (dB), SDR (dB), and %EER on the simulated RoboVoices dataset concerning various SNR (dB) with and without knowledge distillation loss. Because GradSE works with Mel spectrogram features for enhancement, it is impossible to compute SIR and SDR using Mel spectrograms. J. optim. in the table refers to joint optimization. The average confidence interval is 0.1.

		5			10			20		
	Pre-pro./SNR	**SIR**	**SDR**	**EER**	**SIR**	**SDR**	**EER**	**SIR**	**SDR**	**EER**
	Dry clean speech	—	—	5.6	—	—	5.6	—	—	5.6
	Reverb. clean speech	31.5	6.7	7.5	31.5	6.7	7.5	31.5	6.7	7.5
	Oracle Rank-1 MWF	22.7	6.0	5.9	24.7	7.2	5.9	28.2	7.5	5.9
	Unprocessed (noisy)	12.9	0.8	11.2	15.1	2.0	9.2	17.4	5.0	7.8
	FaSNet	17.0	4.8	12.4	19.7	5.4	10.5	24.6	5.8	8.0
	GradSE	—	—	10.2	—	—	8.5	—	—	7.2
J. optim.	FaSNet	21.1	5.0	10.1	22.8	5.9	8.4	26.2	6.3	7.5
J. optim.	GradSE	—	—	9.8	—	—	8.0	—	—	7.1
J. optim.	FaSNet + KD loss	21.8	5.4	9.9	23.0	6.0	8.0	27.3	6.2	7.1
J. optim.	GradSE + KD loss	—	—	9.2	—	—	7.7	—	—	6.8

The results show that adding KD loss to the joint optimization approach further improves the performance in terms of EER. Specifically, the EER values are even lower than the joint optimization approach without KD loss on all SNR levels. This indicates that KD loss is effective in removing the noise and reverberation for speaker verification.

6.4.3 Impact on SV-dedicated dataset

Table 6.4 presents the average EER on different enrollment conditions on the MultiSV dataset (refer to section 3.2.1) using speech enhancement and speaker verification systems. The ReSNet-based speaker verification system is implemented from [Mošner et al., 2022] for comparing the performance. CE stands for clean reverberated enrollment, and MRE stands for multi-channel enrollment.

The oracle Rank-1 MWF achieved an EER of 2.7% on clean enrollment (CE) against unprocessed EER of 4.0%, 1.6% on MRE (multi-channel re-transmitted enrollment) against the unprocessed EER of 5.8%, and 3.1% on the MRE-hard against unprocessed EER of 10.2%. There is some reverberation in the clean enrollment, which explains the improvement in terms of EER. Moreover, there is a great mismatch between the enrollment and test segments among all the conditions. Although MRE seems easier, it is worth mentioning that MRE data comprises recordings from multiple microphones, including low-quality ones.

Noticeably, the jointly optimized models have consistently improved the performance across all the enrollment conditions. We observe from Table 6.4 that Diff-Filter achieved the best performance in all the enrollment conditions. In the case of MRE-hard, which has hard background noise sources including reverberation, diff-Filter improved the speaker verification performance by almost 6% over the unprocessed signal. The jointly optimized diff-Filter outperforms all the models across the enrollment conditions.

Table 6.4: % EER on the MultSV Eval dataset. Mask and filter in the table refer to the speech enhancement output, which is further given as input to the speaker verification. The output of GradSE is the Mel spectrogram. Joint optim. in the table refers to the joint optimization of speech enhancement and speaker verification module. The confidence interval is 0.1.

	SE	SV	CE	MRE	MRE-Hard
	Unprocessed	ECAPA-TDNN	4.0	5.8	10.2
	Oracle Rank-1 MWF	ECAPA-TDNN	2.7	1.6	3.1
	Mask	ReSNet	4.2	3.9	5.3
	Conv-TasNet	ReSNet	4.2	3.7	4.6
	Conv-TasNet	ECAPA-TDNN	3.8	3.6	4.5
	Diff-Estimator	ECAPA-TDNN	3.8	3.6	4.5
	GradSE	ECAPA-TDNN	3.8	3.7	4.5
	Diff-TasNet	ECAPA-TDNN	3.6	3.5	4.5
	Diff-Filter	ECAPA-TDNN	3.6	3.5	4.3
Joint optim.	Conv-TasNet	ECAPA-TDNN	3.77	3.61	4.49
Joint optim.	Diff-Estimator	ECAPA-TDNN	3.81	3.58	4.51
Joint optim.	GradSE	ECAPA-TDNN	3.71	3.64	4.43
Joint optim.	Diff-TasNet	ECAPA-TDNN	3.57	3.55	4.34
Joint optim.	Diff-Filter	ECAPA-TDNN	**3.51**	**3.24**	**4.26**

Table 6.5: % EER on different noise conditions of the VOiCES Eval dataset. The confidence interval is 0.2.

		Noise conditions			
	Testing environment	Clean	Babble	TV	Music
	Unprocessed	4.1	8.8	7.8	7.9
	Oracle Pytorch Rank-1 MWF	2.8	3.9	4.1	4.2
	FaSNet	4.4	7.8	7.4	7.9
	FaSNet Rank-1 MWF WPE	4.2	6.9	6.4	6.8
	GradSE	3.9	6.7	6.2	6.6
	Mask Estimator	4.1	7.1	7.2	7.2
	Diff-Estimator	3.9	6.6	6.6	6.8
	Conv-TasNet	3.9	6.8	6.3	6.4
	Diff-TasNet	3.9	6.3	6.1	6.2
	Diff-Filter	3.8	6.2	6.0	6.0
Joint optimization	FaSNet ECAPA-TDNN	4.0	6.6	6.2	6.5
Joint optimization	GradSE ECAPA-TDNN	3.8	6.4	6.0	6.2
Joint optimization	FaSNet + ECAPA-TDNN + KD loss	3.9	6.6	6.1	6.3
Joint optimization	GradSE + ECAPA-TDNN + KD loss	3.8	6.2	5.9	6.1
Joint optimization	MaskEstimator + ECAPA-TDNN	3.9	6.5	6.2	6.4
Joint optimization	Diff-Estimator + ECAPA-TDNN	3.8	6.4	6.2	6.2
Joint optimization	Conv-TasNet + ECAPA-TDNN	3.8	6.4	6.2	6.3
Joint optimization	Diff-TasNet + ECAPA-TDNN	3.7	6.1	5.7	5.9
Joint optimization	Diff-Filter + ECAPA-TDNN	**3.7**	**5.9**	**5.6**	**5.7**

6.4.4 Validation on a public dataset

We evaluated our proposed approaches also on the publicly available VOiCES eval dataset (refer to section 3.2.1). VOiCES provides speech data in real recording environments, including distant microphones, background noise, and reverberant room acoustics. The data for the eval set was recorded in challenging rooms with different types of microphones. Moreover, the noise source was closer to the microphones.

Table 6.5 shows the results of the joint optimized systems with and without KD loss on the VOiCES dataset. We reported the results on all the distractor noise conditions, namely clean, babble, TV and music. Presumably, the condition without any noise, i.e. clean, has the best EER. The Oracle Rank-1 MWF achieved the best performance, which we considered as the target EER. With a comparatively higher EER, Babble seems to be the hardest among all the noise conditions, with an EER of 8.8% without any pre-processing. However, joint-optimized models have improved the performance. Diff-Filter dropped the EER of Babble to 5.9%. The results show that DPM-based systems are superior to FaSNet-based systems. There is a consistent improvement in the DPM-based systems across all the noise conditions. Although, like all the experiments so far, FaSNet doesn't improve much. When used with KD loss, the beamforming-based system enhances the performance by a margin. Although we didn't train the proposed systems on the VOiCES dataset, it achieved superior performance in all the noise conditions. This shows that our systems are generalized to unseen real recorded data.

6.5 Conclusion

In this chapter, we described the problem of training mismatch that often leads to performance degradation in speaker verification. Training the speech enhancement model independently can lead to a mismatch between the front-end speech enhancement and the back-end speaker verification model. A joint optimization approach of speech enhancement and speaker verification was proposed to solve the mismatch issue. The speech enhancement and speaker verification models trained together consider the interdependencies between the two tasks, leading to the enhancement of feature representation used for speaker verification and improved performance. Similarly, the speaker verification model may be trained to be more robust to noise or other distortions.

We described all the jointly optimized systems' architecture in section 6.2. The joint optimization contains a speech enhancement model and a speaker verification model, mainly based on two approaches. The first approach is based on a beamforming network trained using SI-SDR and cross-entropy loss function. The second approach is based on several diffusion probabilistic model-based speech enhancement techniques, which we have explained in detail in the Chapter (chapter 5). We analyzed the impact of using a novel loss function: similarity-preserving knowledge distillation loss. The similarity-preserving knowledge distillation loss helps the jointly optimized system generate embeddings more similar to those caused by clean speech.

We conducted the experiments on the simulated RoboVoices dataset to evaluate the robustness of the proposed approach concerning different SNR levels. We compared the speech enhancement and speaker verification performance in terms of the respective evaluation metrics, namely SIR, SDR and EER. Usually, speech enhancement evaluation is often ignored, but it is essential to understand how to tune the front end (speech enhancement).

We also computed results using the newly released MultiSV dataset for training and evaluating text-independent multi-channel speaker verification systems. The dataset can be used for experiments with dereverberation, denoising, and speech enhancement. There are four enrollment conditions, but we computed results only on three of them, as the fourth deals with single-channel data, which is out of this thesis's scope.

We also evaluated our systems' effectiveness on the VOiCES dataset, which has real-recorded data in an acoustically challenging environment. Although the models are not trained on VOiCES, the proposed approaches achieved the best results on all the noise conditions showing the generalizability to unseen real-recorded data.

Chapter 7

Conclusion

Contents

This chapter summarizes the contributions made during the thesis work and briefly describes the future work. The primary goal of this PhD thesis was to explore the usage of DNN-based speech enhancement techniques for a mobile security robot in the context of far-field speaker verification under the framework of the project ANR Robovox. In the first step, we proposed to develop algorithms to simultaneously process noise and reverberation inspired by recent works in the speech enhancement domain. The final goal was to propose end-to-end approaches that perform speaker verification directly from multi-channel perturbed signals. The contribution of all the work developed during the thesis is presented in section 7.1, and section 7.2 discusses the future direction of this work.

7.1 Contribution Of Thesis

In chapter 2, we presented the literature review covering signal processing techniques and deep neural network architectures for speaker verification and speech enhancement. We briefly analyzed some of the feature extraction processes for speaker verification. We also described a few factors responsible for performance degradation in various acoustic conditions, especially noise interference and reverberation.

In chapter 3, we presented the evaluation metrics for speech enhancement and speaker verification systems. These metrics assessed the quality of speech signals affected by noise, reverberation, and other distortions. We described the evaluation protocol we followed to test the models in different acoustic scenarios and compare their performance. This chapter also lists publicly available speech enhancement and speaker verification datasets. We introduced RoboVoices, a synthetic dataset explicitly designed for multi-channel speech enhancement. RoboVoices simulates real room environments with additive noise and reverberation from dry speech segments. Designing such a dataset is necessary as training speech enhancement approaches require ground-truth knowledge about the target speech and, to some extent, the degradation. This information is not available in the available corpora for far-field SV.

As a first step towards tackling the noisy and reverberant interferences in far-field

speaker verification, we benchmarked multi-channel speech enhancement techniques used as a pre-processing pipeline in chapter 4. The experiments involved using DNN and a combination of DNN with signal processing methods as a front-end to the x-vector speaker verification system. We developed a multichannel pre-processing pipeline by combining a time-domain neural beamformer (FaSNet), multi-channel Wiener filter (MWF), and weighted prediction error (WPE). We analyzed the pre-processing pipeline's impact in scenarios with additive noise and reverberation on speaker verification performance. We have also studied the importance of acoustic conditions during the enrollment phase and found that the integrated DNN and signal processing approach showed robustness to low SNR scenarios. The results from experiments on synthetic and VOiCES datasets showed that combining DNN with signal processing improved speaker verification performance and that matching acoustic conditions during enrollment and testing reduced the impact of reverberation. The proposed approach achieved the best performance across noise conditions on the VOiCES dataset, demonstrating its generalization to real recorded data as the model was trained on synthetic data.

We dedicated chapter 5 to exploring different diffusion-based multi-channel speech enhancement techniques to tackle the challenging task of strong environmental noise and reverberation in a far-field speaker verification system. Our contributions are mainly for front-end processing. We proposed a novel technique for computing the time-frequency masks and multi-channel filters and changed the architecture using scoring-based diffusion models. We studied the impact of utterance length on speaker verification and showed that shorter lengths severely impacted speaker verification performance. The evaluation was done using speech enhancement metrics (SDR, SIR) and EER, which showed that diffusion-based techniques improved performance against the state-of-the-art baselines and the Diff-Filter approach achieved the best performance while maintaining signal quality. Using the speech enhancement model as a pre-processing improves the results, as observed in the speaker verification performance with the unprocessed signal.

In chapter 6, a joint optimization approach for speech enhancement and speaker verification was proposed to solve the problem of training mismatch that leads to performance degradation in speaker verification (SV). The joint optimization system consists of speech enhancement and SV models. We proposed two joint optimization approaches, the FaSNet-based approach and diffusion probabilistic model-based speech enhancement techniques. The impact of a novel loss function, the similarity-preserving knowledge distillation loss, was also analyzed. Experiments were conducted on the simulated RoboVoices dataset, a newly designed dataset for text-independent multi-channel speaker verification, and the VOiCES dataset with real-recorded data in challenging acoustic environments. The joint optimization framework provides better performance than individual or separately trained models.

7.2 Future Direction

The future work can be envisioned focusing on three aspects: (i) improving the system performance through modifications in network architectures, (ii) using novel fine-tuning training procedures based on self-supervised learning, and (iii) making proposed architectures of speech enhancement and speaker verification suitable for deployment on embedded devices.

In chapter 4, we implemented WPE as a dereverberation technique to reduce the artefacts introduced through room geometry. Recently, DNN-enabled dereverberation tech-

niques have been proposed and significantly improved over the classical WPE algorithm. In addition, DNN-based dereverberation techniques will allow the joint optimization of speech enhancement, dereverberation and speaker verification as a single network architecture.

Self-supervised learning can utilize unlabeled data to learn useful representations, which has recently received more and more attention in many fields. Many of these approaches propose the usage of the same speech utterance with and without data augmentation, given to speaker verification embedding extractor and the network is trained to maximize the similarity scores. We conducted preliminary experimentation with self-supervised learning, which has shown promising results. Therefore, it will be an imperative step to replicate the success of self-supervised learning to conduct similar experimentation with a jointly optimized robust speech enhancement and speaker verification pipeline. Thus, the usage of self-supervised learning will enable the usage of large-scale unlabeled noisy multichannel speech signals for which target clean speech data is not available.

In this thesis work, we presented knowledge distillation-based joint optimization of FaSNet and GradSE speech enhancement systems. It will be interesting to carry forward knowledge distillation-based training on other proposed speech enhancement systems such as Diff-TasNet, Diff-Filter and Diff-Estimator. Furthermore, using large-scale speech language models such as WavLM and UniSpeech-SAT can provide better latent representation space for the speaker as a teacher network during the knowledge distillation phase.

In addition, it's necessary to adapt the deep learning techniques used in the context of model deployment in real-world conditions of on-device scenarios. The deep learning-enabled embedded systems have inherent constraints of memory, and flash memory, on top of latency limits. Therefore, it will be a natural step to consider various deep neural network model optimization approaches, such as pruning and quantization of speech enhancement and speaker verification models. Since the last few years, knowledge distillation has been popularly used for creating lighter DNN network architecture by distillation of knowledge from large DNN pre-trained models for both speech enhancement and speaker verification by keeping the system performance at bay. Lastly, it is also essential in making PyTorch-based DNN models prepare for cross-platform seamless deployment across embedded platforms for which the ONNX library can be a suitable solution.

Appendix A

French Thesis Summary

Contents

A.1 Introduction

La parole est un outil de communication efficace qui transmet de nombreuses informations, y compris les émotions et l'identité du locuteur. Les progrès technologiques ont permis aux machines de converser avec les humains, ce qui a conduit au développement d'applications vocales et à l'essor des assistants vocaux. Les outils basés sur la parole, tels que la reconnaissance vocale et la synthèse vocale, sont essentiels pour l'interaction homme-machine, et les outils pour la sécurité, telles que la vérification du locuteur et les technologies anti-spoofing, sont cruciales pour l'authentification des utilisateurs. Les technologies vocales sont confrontées à un défi de taille dans les environnements bruyants, qui déforment considérablement le signal vocal, comme illustré dans la figure 1. En outre, les signaux vocaux sont réfléchis par divers objets présents dans l'environnement, ce qui crée une réverbération. Le bruit environnant et la réverbération dégradent la qualité et l'intelligibilité de la parole, ce qui a un impact négatif sur les applications vocales telles que la vérification du locuteur. Si les êtres humains sont capables de pallier les interférences dans les communications, les machines ne sont pas encore aussi avancées.

Pour relever les défis posés par les environnements bruyants, les technologies vocales nécessitent des algorithmes avancés de traitement de la parole afin d'améliorer la qualité et l'intelligibilité de la parole. Le rehaussement de la parole peut être utilisé pour améliorer la qualité perceptive de la parole en estimant des signaux de parole propres à partir des signaux affectés par le bruit et la réverbération. Ceci est particulièrement important pour les applications basées vocales, y compris la vérification du locuteur, qui peut être améliorée par l'ajout d'un prétraitement de rehaussement de la parole. Malgré les recherches en cours dans le domaine du rehaussement de la parole, cela reste un défi dans les scénarios avec une prise de son distante. La figure 2 illustre le processus générique d'amélioration de la parole.

L'objectif de cette thèse est de développer des techniques de rehaussement de parole multicanal pour un robot de sécurité mobile et autonome visant à surveiller la présence humaine dans des locaux industriels pendant les périodes d'inactivité. La vérification du locuteur dans le contexte d'un robot de sécurité mobile fait face à plusieurs défis liés à la vérification de l'identité d'une personne avec une prise de son distante dans des conditions réelles qui peuvent réduire considérablement les performances, comme le bruit ambiant, la réverbération, l'impact des énoncés courts dans la vérification du locuteur, et une inadéquation entre l'enrôlement et le test. Cette thèse vise à proposer et à développer des techniques de rehaussement de parole multicanal basées sur le traitement du signal et les réseaux neuronaux profonds pour pour éliminer le bruit ambiant et la réverbération pour la vérification du locuteur avec prise de son distante.

Nous avons développé diverses approches de rehaussement de la parole multicanal pour la vérification du locuteur avec prise de son distante. Les principales contributions de ce travail comprennent la proposition et la mise en œuvre d'une approche de référence

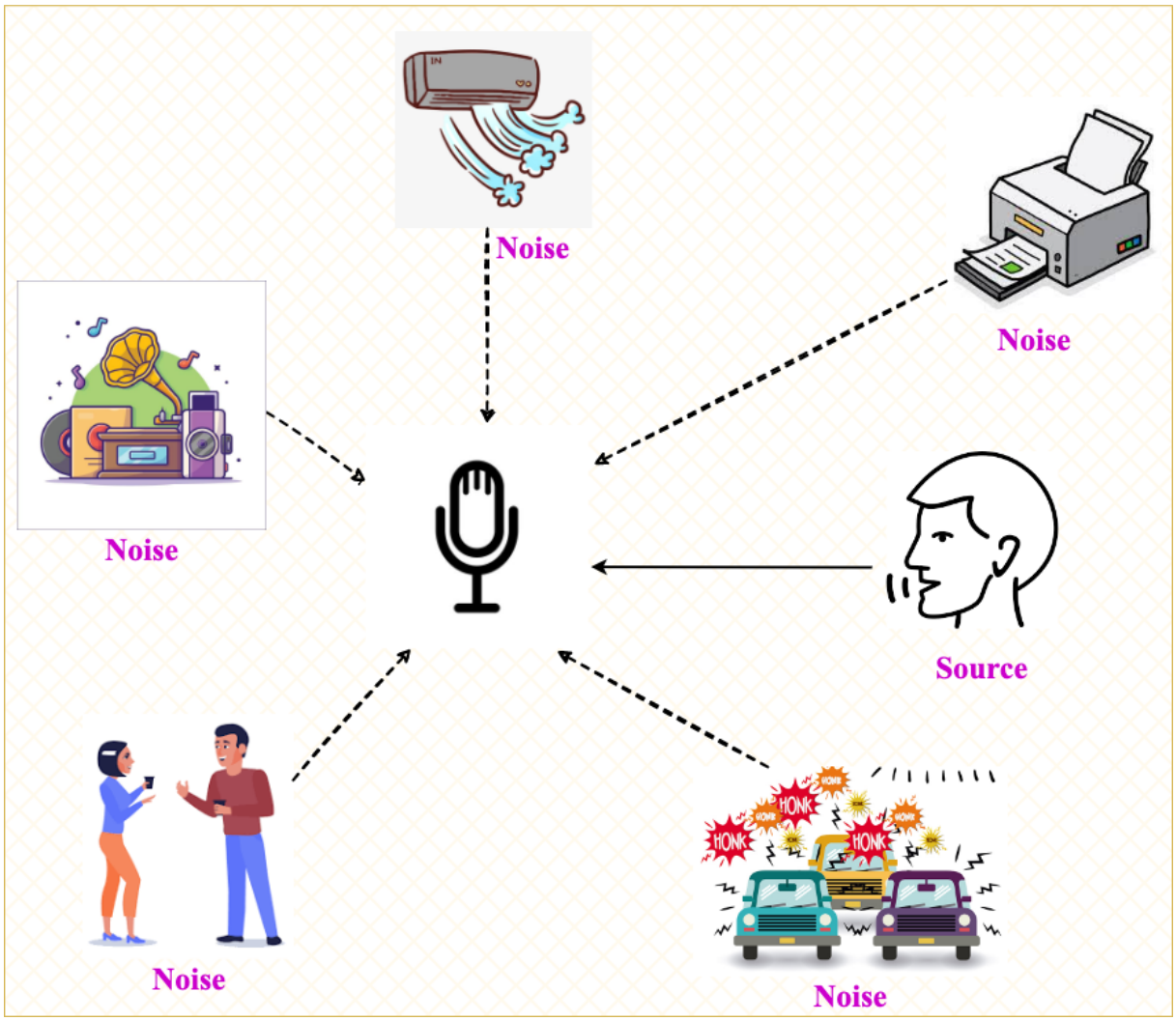

Figure A.1: Illustration d'un environnement bruyant.

de rehaussement la parole multicanal, d'une approche générative basée sur les modèles probabilistes de diffusion, et l'optimisation conjointe des modules de rehaussement de parole et de vérification du locuteur. L'étude examine également l'impact des approches de rehaussement de parole et de vérification du locuteur dans divers scénarios acoustiques, le traitement des données d'enrôlement, la robustesse à un faible rapport signal à bruit, la généralisation à des données non vues et à des énoncés de courtes durées. Les solutions proposées présentent une meilleure robustesse au bruit et sont plus performantes que les techniques de prétraitement de l'état de l'art.

Nous donnons également un aperçu des systèmes de vérification du locuteur et des approches de rehaussement de parole. Nous discutons de la mise en œuvre de diverses architectures pour la vérification du locuteur et le rehaussement de parole, ainsi que de leurs limites dans des scénarios avec prise de son distante. L'importance de l'extraction des coefficients dans les systèmes de vérification du locuteur est soulignée, et nous discutons également de l'impact des interférences sonores sur les performances du système. Nous analysons diverses approches de rehaussement de parole proposées pour améliorer les performances de la vérification du locuteur dans des scénarios avec prise de son distante, impliquant réverbération et bruit d'environnement, ce qui est l'objectif principal de ce travail de thèse.

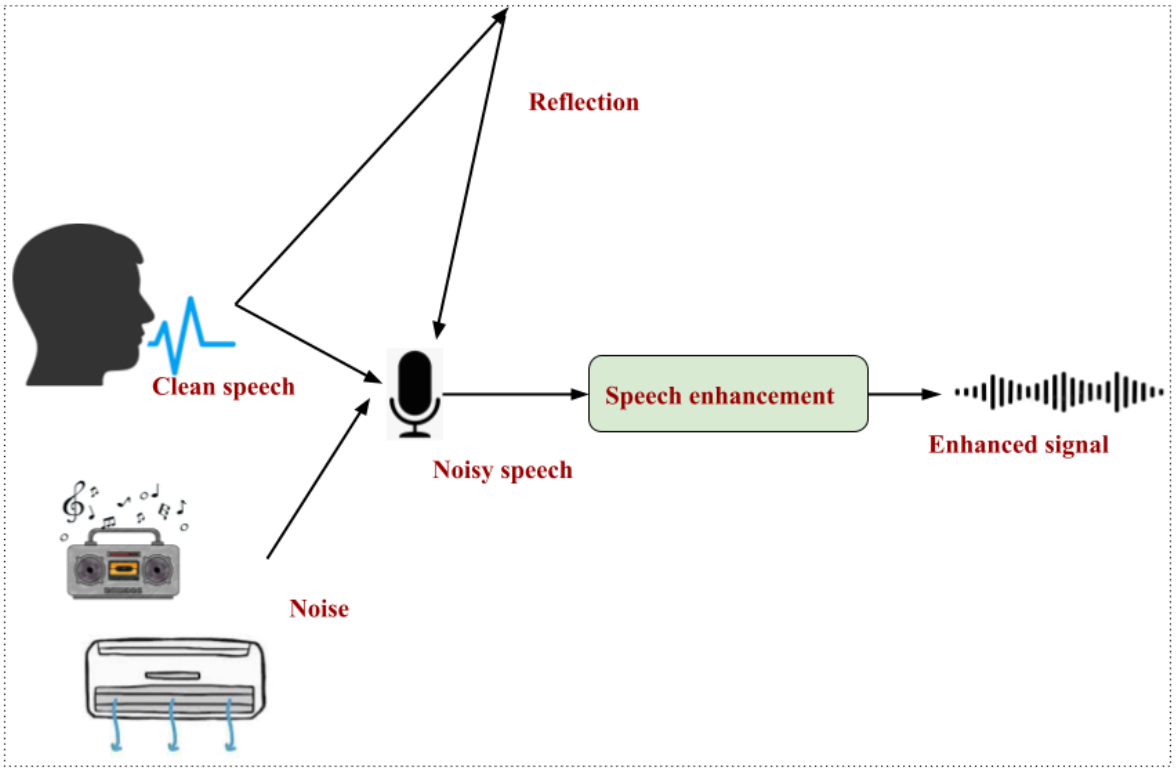

Figure A.2: A graphical representation of generic speech enhancement process.

A.2 Etat de l'art

L'état de l'art en matière de vérification de locuteur évolue constamment à mesure que de nouvelles recherches et avancées technologiques émergent. Les modèles de mélange gaussien ont été une approche populaire en matière de vérification de locuteur pendant de nombreuses années. Ils modélisent la distribution de probabilité des caractéristiques de la parole et comparent la vraisemblance de différents locuteurs en fonction des modèles. Les approches récentes basées sur les DNN ont montré un grand succès en matière de vérification de locuteur en extrayant des caractéristiques de haut niveau à partir de signaux de parole. Ils peuvent être entraînés à apprendre des plongements de locuteur discriminatoires, qui sont utilisés pour comparer et vérifier l'identité du locuteur. En plus de ces approches, il existe également diverses techniques utilisées pour améliorer les performances de vérification de locuteur, telles que l'augmentation de données, l'adaptation de domaine et la fusion de plusieurs systèmes. Dans l'ensemble, l'état de l'art en matière de vérification de locuteur évolue continuellement et devrait continuer à s'améliorer avec le développement de nouvelles technologies et recherches.

L'amélioration de la parole fait référence au processus d'amélioration des signaux de parole qui ont été dégradés par différents types de bruits et de distorsions. Il existe plusieurs approches de pointe qui ont montré des résultats prometteurs en matière d'amélioration de la parole. Les DNN ont été appliqués avec succès à l'amélioration de la parole en apprenant à mapper les signaux de parole bruyants sur des signaux de parole propres. Ils peuvent être entraînés en utilisant différentes fonctions de perte, telles que l'erreur quadratique moyenne, pour minimiser la différence entre les signaux de parole prédits et propres.

Nous présentons un aperçu des systèmes de vérification de locuteur et des approches d'amélioration de la parole au fil des ans pour les deux tâches. Nous avons discuté de la mise en œuvre de diverses architectures pour la vérification de locuteur et l'amélioration de la parole et de leurs limites dans des scénarios éloignés. L'importance de l'extraction de caractéristiques dans le système de vérification de locuteur est soulignée, et nous avons également discuté de l'impact de l'interférence de bruit sur les performances du système. Nous avons analysé diverses approches d'amélioration de la parole qui ont été proposées pour améliorer les performances de vérification de locuteur dans des scénarios éloignés bruyants et réverbérants, qui est l'objectif principal de ce travail de thèse.

A.3 Données et métriques d'évaluation

De grandes quantités de données de parole sont indispensables pour l'apprentissage des modèles, et leur évaluation repose sur différentes métriques. Nous présentons ici les ensembles de données et les métriques d'évaluation utilisés dans ce travail de thèse.

A.3.1 Données

Nous présentons ici les ensembles de données de vérification du locuteur que nous avons utilisé pour nos expérimentations. L'ensemble de données pour la vérification du locuteur doit être diversifié et équilibré, et comprendre des enregistrements de nombreux locuteurs dans des conditions différentes. Les enregistrements audio peuvent être dépendants ou non du texte. Dans le cas d'un système indépendant du texte, le locuteur parle librement sans aucune contrainte particulière. Il est essentiel de tenir compte de l'application cible lors de la conception du système et de choisir les données d'entraînement et de test en conséquence. Les données utilisées peuvent être divisées en plusieurs ensembles : apprentissage, développement et évaluation ; l'ensemble d'évaluation peut être composé de plusieurs parties permettant des évaluations d'aspects spécifiques.

La création d'un système de rehaussement de la parole basé sur un DNN performant nécessite une grande quantité de données d'apprentissage, qui sont coûteuses et difficiles à collecter. Pour être efficaces, les ensembles de données pour le rehaussement de la parole doivent être diversifiés, de haute qualité et de grande taille. Nous décrivons quelques ensembles de données couramment utilisés pour le rehaussement de parole multicanal. La connaissance des signaux de référence (parole propre) et des signaux bruitées est nécessaire pour l'apprentissage des approches de rehaussement de la parole ; et il y a un manque d'ensembles de données d'apprentissage multicanal pour la vérification du locuteur. Nous avons donc fabriqué un ensemble de données synthétiques appelé RoboVoices pour répondre à la fois aux exigences de vérification du locuteur et de rehaussement de parole multicanal.

L'ensemble de données RoboVoices a été créé en simulant la réponse impulsionnelle de la pièce (RIR – Room Impulse Response) et en l'appliquant sur de la parole propre, et du bruit. L'ensemble de données est divisé en deux parties : ensemble d'apprentissage, et ensemble d'évaluation. Les données vocales proviennent du sous-ensemble propre de Librispeech et du corpus Fabiole, avec 25 heures pour l'apprentissage et 2 heures pour l'évaluation. L'ensemble de données de bruit a été préparé en extrayant des bruits réalistes à partir de Freesound et en les divisant en deux ensembles, l'un pour l'apprentissage, et l'autre pour l'évaluation. Nous avons créé deux ensembles d'évaluation, RoboVoices eval1 et RoboVoices eval2. RoboVoices eval1 a été créé à partir de la parole propre de Fabiole et

les bruits extraits de freesound ; et RoboVoices eval2 a été créé à partir de la parole propre de Fabiole et les bruits extraits de MUSAN. Le fait de disposer de plusieurs ensembles d'évaluation permet d'évaluer les performances des systèmes dans des conditions de bruit différentes.

Nous avons utilisé la boîte à outils pyroomacoustics pour générer un corpus de réponses impulsionnelles de salle (RIR). Nous avons créé 10000 RIR pour l'apprentissage et 3600 pour l'évaluation, avec les dimensions de la pièce et les coefficients d'absorption sélectionnés au hasard pour obtenir une valeur RT60 souhaitée. Le bruit et la parole propre ont été convolués avec la réponse impulsionnelle de salle pour simuler la parole et le bruit réverbérés ; et ont été additionnées, en respectant différents niveaux de rapport signal à bruit pour fabriquer les ensembles d'apprentissage et d'évaluation. Le SNR est tiré au hasard avec une distribution uniforme entre 0 dB et 10 dB pour l'ensemble d'apprentissage et sélectionné à 5 dB, 10 dB ou 20 dB pour l'ensemble d'évaluation. Au total, 10000 mélanges pour la formation et 3600 mélanges pour l'évaluation sont générés à l'aide de l'ensemble de données Fabiole.

A.3.2 Métriques d'évaluation

Il existe deux approches pour évaluer les modèles d'amélioration de la parole: objective et subjective. L'évaluation objective implique le calcul de la qualité du signal de sortie, tandis que l'évaluation subjective exige que des humains évaluent la sortie. Nous avons utilisé l'approche objective pour évaluer les modèles de rehaussement de parole et de vérification du locuteur utilisés dans cette thèse. Nous avons utilisé le rapport signal à distorsion (SDR – Signal to Distortion Ratio) et le rapport signal-à interférence (SIR – Signal to Interference Ratio) pour évaluer les systèmes de rehaussement de la parole et le taux d'égale erreur (EER – Equal Error RAte) pour évaluer les systèmes de vérification du locuteur.

L'évaluation a consisté à traiter des signaux correspondant à différentes conditions acoustiques : parole bruitée (mélange de bruit et de parole propre réverbérée) avec ou sans application du rehaussement de parole, parole propre non réverbérée et parole propre réverbérée. Selon les cas, les évaluations ont impliqué le calcul des métriques SDR, SIR et/ou EER, sur les corpus RoboVoices, MultiSV et VOiCES.

A.3.3 Rehaussement de parole multicanal pour la vérification du locuteur avec prise de son distante

Nous avons proposé de rehausser la parole pour la vérification du locuteur avec prise de son distante. Les approches proposées reposent soit un filtrage basé sur un DNN, soit une combinaison de DNN et d'approches de traitement du signal. FaSNet est une technique neuronale de formation de voies, pour calculer les masques temps-fréquence qui intègrent les informations de phase. La seconde approche repose sur la combinaison de FaSNet avec un filtre de Wiener multicanal de rang 1 (Rank-1 MWF – Multicanal Wiener Filter) et la prise en compte de l'erreur de prédiction pondérée (WPE – Weighted Prediction Error) pour la déréverbération. Le chapitre correspondant (Chapitre 4) évalue également l'impact de ces approches dans divers scénarios acoustiques bruyants et réverbérés à différents niveaux de rapport signal à bruit. Ces techniques servent de référence pour le rehaussement de parole multicanal appliqué à la vérification du locuteur avec prise de son distante.

A.3.4 Rehaussement de parole multicanal par DNN

Le pipeline de rehaussement de la parole proposé pour la vérification du locuteur commence par le module neuronal FaSNet de formation de voies pour réaliser une première séparation de la parole et du bruit. Les estimations de la parole et du bruit sont alors utilisées pour calculer les matrices de covariance de la parole et du bruit, qui servent à calculer le filtre MWF de rang 1. La variante pondérée de la distorsion de la parole (SDW – Speech Distortion Weighted) du filtre MWF de rang 1 est utilisée pour estimer le filtre. La sortie du filtre Rank-1 SDW-MWF est traitée avec WPE pour atténuer la réverbération. Le pipeline est décrit dans la figure 3.

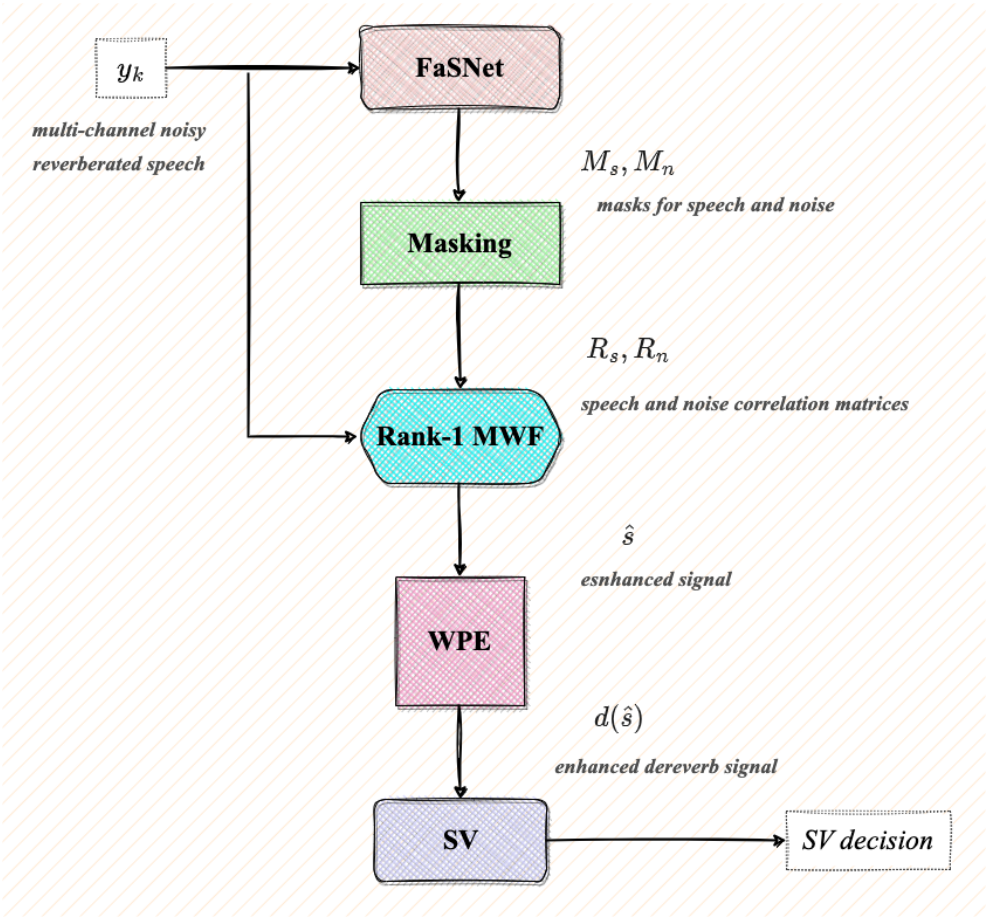

Figure A.3: Représentation graphique du pipeline de prétraitement proposé et utilisé dans nos expériences. Le traitement prend en entrée un signal y_k multicanal bruité et réverbéré. La sortie de FaSNet sert pour estimer les masques M_s et M_n de la parole et du bruit. Ces estimations sont utilisées pour calculer les matrices de covariance R_s et R_n pour la parole et le bruit, qui sont à leur tour, utilisées pour calculer le filtre MWF de rang 1. Nous appliquons alors le module WPE pour réduire la réverbération. Le signal de de parole rehaussé et déréverbéré $d(\hat{s})$ est donné en entrée au système de vérification du locuteur.

FaSNet repose sur un traitement en deux étapes : la première étape calcule un filtre de formation de voies pour un canal de référence, et la deuxième étape estime les filtres de formation de voies pour les canaux restants en utilisant le filtre de sortie de la première étape.

L'entrée de chacune des deux étapes comprend le canal cible et la corrélation croisée normalisée (NCC – Normalized Cross Correlation) entre les canaux. La NCC mesure la similarité entre les signaux et est calculée au niveau de la trame. Cela permet à FaSNet d'intégrer des informations sur les corrélations entre les canaux, ce qui conduit à une meilleure estimation du filtre et à une amélioration de la qualité du signal estimé. FaSNet utilise des réseaux temporels convolutifs pour un traitement à faible latence et est appris à l'aide des fonctions de coût SI-SDR et SI-SNR en fonction de la tâche à résoudre.

Nous avons utilisé les techniques de réduction du bruit basées sur le filtre de Wiener multicanal pondéré par la distorsion de la parole (SDW-MWF), qui nécessite l'estimation des matrices de corrélation de la parole et du bruit. FaSNet est utilisé pour estimer les masques temps-fréquence pour la parole et le bruit, et le masque de parole prédit est appliqué au mélange pour obtenir une estimation du signal de parole. Les matrices de corrélation spatiale requises pour le filtrage SDW-MWF sont calculées à l'aide des estimations de la parole et du bruit.

L'algorithme de déréverbération, qui repose sur l'erreur de prédiction pondérée (WPE), vise à réduire les composantes réverbérantes dans le signal vocal acquis tout en préservant la composante directe du signal. WPE estime le filtrage optimal au sens du maximum de vraisemblance en utilisant le signal vocal réverbéré observé.

Le modèle FaSNet exploite des blocs DPRNN (Dual Path RNN) au lieu des blocs TCN (Temporal Convolutive Network) pour la formation de voies. Les sorties séparées de parole et de bruit de FaSNet sont utilisées pour calculer les masques cibles. Nous avons utilisé µ comme facteur de compromis entre l'interférence et la distorsion dans le filtre MWF de rang 1 avec un masque idéal (IRM – Ideal Ratio Mask) et l'avons fait varier de 0,1 à 0,9. Après expérimentation, il s'est avéré qu'une valeur µ de 0,1 était l'hyperparamètre optimal pour améliorer la performance généralisée dans toutes les conditions de rapport signal à bruit. Nous avons également constaté que plus la valeur de µ est faible, meilleure est la performance de la vérification du locuteur. Le paramètre µ du SDW-MWF a donc été fixé à 0,1 pour limiter la distorsion introduite par le filtre.

Nous avons également utilisé, à titre de comparaison, les formations de voies basées sur le masquage pour le rehaussement de parole multicanal décrits dans. Les auteurs ont proposé d'utiliser l'approximation de rang-1 pour estimer les matrices de covariance de la parole pour le beamformer MVDR (Minimum Variance Distortionless Response) et ont constaté que les approches GEVBAN (Generalized Eigenvalue beamformer with Blind Analytic Normalizaation) et Rank-1 MVDR étaient plus performantes que l'approche Rank-1 MWF. Pour l'estimation de l'IRM, ils ont utilisé des cellules BLSTM (Bidirectional Long Short Term Memory) et MSE (Mean Square Error) comme fonction de coût. Pour les expériences de vérification du locuteur, nous avons utilisé les outils Kaldi pour l'extraction des x-vectors ; le DNN a été appris avec des données augmentées du corpus Musan. Les coefficients MFCC (Mel Frequency Cepstral Coefficients) ont été normalisées et les trames correspondant à du silence ont été supprimées grâce à une VAD (Voice Activity Detection). Le système a été évalué à l'aide du corpus Fabiole, avec 3441 fichiers utilisés pour l'enrôlement et les fichiers restants ont servi pour le test. Le classifieur PLDA a été entraîné sur 200k x-vecteurs provenant de VoxCeleb qui ont été réduits à une dimension de128 grâce à une analyse discriminante linéaire. Les expériences ont été menées, d'une part, dans des conditions de SNR similaires entre enrôlement et test, et d'autre part, avec un enrôlement correspondant à des conditions SNR multiples. Des évaluations ont également été menées en appliquant les approches de rehaussement de parole sur les données d'enrôlement. L'application des mêmes pré-traitements sur les données

Table A.1: Résultats sur les ensembles de données RoboVoices eval1 et RoboVoices eval2 en utilisant différentes méthodes de prétraitement. RoboVoices eval1 est composé de parole propre de Fabiole et de bruit de Freesound. RoboVoices eval2 est composé de parole propre provenant de Fabiole et de bruits provenant de MUSAN. L'intervalle de confiance moyen est de $0,1$.

Noise type	**RoboVoices eval1**			**RoboVoices eval2**		
Pre-processing/SNR	**SDR**	**SIR**	**EER**	**SDR**	**SIR**	**EER**
Dry speech	—	—	14.9	—	—	14.9
Reverb-speech	—	—	20.6	—	—	20.6
Unprocessed (Noisy)	2.6	15.1	28.2	2.4	14.8	25.7
BLSTM Rank-1	5.4	20.7	26.8	5.1	20.7	23.1
FaSNet	5.3	20.5	38.7	4.7	18.3	32.6
BLSTM GEV-BAN	5.4	20.6	27.1	4.3	20.2	23.8
FaSNet GEV-BAN	5.8	21,9	26.8	5.6	21.2	22.2
BLSTM MVDR Rank-1	5.8	20.1	27.0	4.9	20.5	23.7
FaSNet Rank-1 MWF	6.1	21.0	24.9	5.9	21.5	21.9
FaSNet Rank-1 MWF WPE	**7.0**	**21.0**	**23.3**	**6.1**	**21.5**	**20.5**

d'enrôlement et de test permet d'éliminer des sources potentielles de variabilité.

A.3.5 Résultats

Nous présentons ici les résultats obtenus avec diverses approches de rehaussement de parole multicanal appliquées en prétraitement sur les ensembles de données RoboVoices eval1 et eval2 pour la vérification du locuteur avec prise de son distante, dans des scénarios bruyants et réverbérés. L'évaluation est faite en termes d'EER, de SIR et de SDR, et l'enrôlement est toujours effectué en utilisant de la parole propre.

A.3.6 Comparison avec les approches de référence

Les résultats du tableau A.1 montrent que les techniques de formation de voies basées sur le rang 1 sont plus performantes que FaSNet en termes d'EER. Toutefois, l'intégration de FaSNet, du filtre MWF de rang 1 et de WPE apporte une amélioration substantielle par rapport à toutes les autres techniques, ce qui montre les avantages de l'utilisation combinée de plusieurs traitements dans des scénarios bruyants et réverbérés.

A.3.7 Impact de l'enrôlement

Le tableau A.2 présente les résultats de vérification du locuteur pour différentes conditions acoustiques et différents pré-traitements appliqués sur les données d'enrôlement et de test. Les résultats, en terme de taux d'égale erreur, sont fournis pour l'ensemble de données RoboVoices eval1.

Les résultats montrent clairement que pour chaque condition de test, les meilleures performances sont obtenues lorsque les données d'enrôlement correspondent aux mêmes conditions acoustiques et au même pré-traitement que les données de test. D'autre part

Table A.2: % EER en fonction des conditions acoustiques et des pré-traitements appliqués sur les données d'enrôlement et sur les données de test. Les évaluations sont faites sur l'ensemble de données RoboVoices eval1. L'intervalle de confiance moyen est de $0,1$.

Test conditions	**Enrollment conditions**				
	Dry speech	Reverb speech	Unprocessed (Noisy)	MVDR Rank-1	FaSNet Rank-1 MWF WPE
Dry speech	**14.9**	15.4	16.7	16.1	15.7
Reverb. speech	20.6	**19.8**	20.5	20.4	20.1
Unprocessed (Noisy)	28.2	24.9	**23.8**	24.9	24,3
MVDR Rank-1	27.0	24.2	23.4	**21.3**	22.5
FaSNet Rank-1 MWF WPE	23.3	22.8	21.5	22.4	**19.2**

sur les données bruitées et réverbérées, c'est le rehaussement de parole avec l'approche combinée FaSNet Rank-1 MWF WPE qui conduit aux meilleures performances.

A.3.8 Impact du rapport signal à bruit

Table A.3 presents the experimental results for speaker verification on the RoboVoices eval1 dataset, evaluated under match SNR and multi-SNR conditions.

Table A.3: % EER sur les données RoboVoices eval1 en fonction du rapport signal à bruit des données de test, et du rapport signal à bruit des données d'enrôlement. L'intervalle de confiance moyen est de $0,1$.

Test conditions	**Enrollment conditions**			
SNR (dB)	**Match SNR**			**Multi-SNR**
	5	10	20	
5	18.6	18.4	18.5	18.7
10	14.4	14.3	13.7	14.4
20	11.8	11.4	11.1	11.3

Le tableau A.3 présente les résultats de vérification du locuteur obtenus sur les données RoboVoices eval1, en fonction des rapport signal à bruit des données d'enrôlement et de test.

Les résultats montrent que plus les données sont bruitées (SNR faible), plus les performances sont mauvaises. Cependant, pour chaque niveau SNR des données de test, le SNR des données d'enrôlement a un impact mineur. L'enrôlement multi-SNR n'apporte pas de gain en performances en comparaison d'un enrôlement avec les données les moins bruitées (i.e., SNR de 20 dB).

A.3.9 Impact des pré-traitements

Le tableau A.4 montre les performances de reconnaissance du locuteur sur l'ensemble de données RoboVoices eval1, pour diverses approches de prétraitement et pour dif-

férentes conditions de rapport signal à bruit. Les approches de prétraitement comprennent BLSTM GEV-BAN, BLSTM Rank-1 MWF, BLSTM MVDR Rank-1, FaSNet, FaSNet GEV-BAN, FaSNet Rank-1 MWF et FaSNet Rank-1 MWF WPE.

Table A.4: % EER sur les données RovoVoices eval1 pour différentes approches de prétraitement (rehaussement de parole). L'intervalle de confiance moyen est de 0,1.

Pre-proces./SNR	**5**	**10**	**20**
Unprocessed (Noisy)	34.4	28.0	22.2
BLSTM GEV-BAN	32.5	26.8	21.9
BLSTM Rank-1 MWF	32.2	26.5	21.6
BLSTM MVDR Rank-1	32.3	26.6	22.1
FaSNet	45.7	39.0	31.5
FaSNet GEV-BAN	32.0	26.3	22.0
FaSNet Rank-1 MWF	29.9	24.0	20.8
FaSNet Rank-1 MWF WPE	**27.1**	**23.2**	**19.7**

Les résultats indiquent une réduction de l'EER pour toutes les approches de prétraitement par rapport à un signal d'entrée non traité, FaSNet Rank-1 MWF WPE étant le plus performant pour toutes les valeurs de SNR et montrant la plus grande robustesse aux conditions de faible SNR avec une réduction absolue de l'EER de 7% à 5 dB. Les résultats suggèrent que l'utilisation combinée de FaSNet et de Rank-1 MWF est particulièrement efficace dans les scénarios à faible SNR.

A.3.10 Impact de la durée de parole

Table A.5: % EER sur les données RoboVoices eval1 pour différentes durées de parole. Pour chaque catégorie (moins ou plus de 4 sec.) les performances indiquées correspondent à une moyenne par rapport aux différents niveaux SNR. L'intervalle de confiance moyen est de 0,1.

	Testing environment	**EER**	**EER**
	Utterance length	**Below 4 secs**	**Above 4 secs**
	Dry clean speech	18.8	5.6
	Reverberated clean speech	21.1	7.5
	Unprocessed (Noisy)	27.8	9.5
	BLSTM Rank-1	27.4	9.2
	BLSTM MVDR Rank-1	27.5	9.4
	FaSNet	28.1	10.3
	FaSNet Rank-1 WPE	27.5	9.0

Le tableau A.5 présente les performances de reconnaissance du locuteur sur les données RoboVoices eval1, en fonction de la durée des énoncés, plus ou moins de 4 secondes. Les résultats montrent une dégradation importante des performances pour les énoncés courts (moins de 4 secondes), en comparaison des performances observées sur les énoncés longs (plus de 4 secondes) ; et ce quels que soient les conditions acoustiques et les

pré-traitements appliqués. Les techniques de base de rehaussement de parole n'ont pas amélioré les performances, et l'approche FaSNet les a même dégradées en raison des artefacts introduits. Cependant, l'approche combinée FaSNet Rank-1 MWF WPE a réussi à améliorer légèrement les performances.

A.3.11 Validation sur un ensemble de données publiques

Table A.6: % EER pour différentes conditions de bruit sur l'ensemble de données VOiCES Eval. L'intervalle de confiance est de 0,2.

SV pre-processing	**Noise conditions**			
	Clean	Babble	TV	Music
Noisy	4.4	9.2	7.9	8.4
BLSTM Rank-1	4.4	8.1	7.1	7.3
BLSTM MVDR Rank-1	4.3	7.3	6.5	6.9
FaSNet	4.4	7.8	7.4	7.9
FaSNet Rank-1 MWF	4.5	7.1	6.8	7.1
FaSNet Rank-1 MWF WPE	**4.0**	**6.3**	**6.0**	**6.4**

Le tableau A.6 présente les performances de reconnaissance du locuteur sur les données d'évaluation VOiCES, pour différentes conditions de bruit. Parmi les 11 positions de microphone de l'ensemble Eval, nous avons sélectionné 3 positions représentatives : la plus proche du locuteur (microphone 2), à mi-distance (microphone 9) et la plus éloignée (microphone 4). Nous avons choisi le microphone le plus proche du locuteur comme microphone de référence pour les traitements par FaSNet.

Les résultats montrent que les meilleures performances de vérification du locuteur sont obtenues pour la condition sans bruit (parole propre), et ce, quel que soit le prétraitement mis en œuvre. Le bruit babble apparait comme la condition la plus difficile pour la reconnaissance du locuteur. Toutes les approches de rehaussement de la parole permettent d'améliorer les performances quel que soit le bruit (babble, TV ou musique).

Le gain le plus important est là encore obtenu avec l'approche combinée de rehaussement de parole FaSNet Rank-1 MWF WPE, et pour laquelle on observe des performances assez similaires pour les trois types de bruits (babble, TV et musique). Cette approche a démontré son efficacité, même avec un apprentissage effectué à partir de données synthétiques générées pour des scénarios spatiaux génériques et éventuellement mal adaptés. L'étude suggère que la conception d'un ensemble d'entraînement avec des conditions adaptées pourrait contribuer à améliorer encore plus les performances sur l'ensemble de données VOiCES.

A.4 Modèles probabilistes de diffusion pour rehaussement de parole multicanal

Nous proposons la mise en œuvre et l'utilisation d'un DPM basé sur les scores pour le rehaussement de la parole multicanal en tant que pré-traitement pour la vérification du locuteur avec prise de son distante. Nous avons mis en œuvre un modèle DPM basé sur les scores qui utilise une équation différentielle stochastique (SDE – Stochastic Differential

Equation) pour la génération d'échantillons et qui diffuse progressivement la distribution des données vers une distribution de bruit. Le modèle génératif basé sur les scores calcule les gradients de la densité de probabilité logarithmique du bruit pour apprendre l'inversion temporelle à l'aide d'une SDE. Cela permet une reconstruction plus souple des données à partir du bruit avec différents paramètres et un contrôle de la qualité du son et de la vitesse d'inférence. Les sections suivantes décrivent les techniques de prétraitement multicanal développées à l'aide de DPM pour la vérification du locuteur.

A.4.1 GradSE

Le modèle GradSE proposé est un pré-traitement de rehaussement de la parole multicanal pour la vérification du locuteur. Il comprend un réseau de conditionnement et un réseau de décodage basé sur le DPM pour définir une distribution de bruit conditionnelle et apprendre les trajectoires du processus de diffusion avant. Le processus de diffusion avant de GradSE utilise une SDE pour ajouter une perturbation du bruit à la distribution des données, et le processus de diffusion arrière est formulé à l'aide d'une SDE inverse. Le réseau neuronal prédit les gradients de la densité logarithmique du bruit, ce qui permet au processus de diffusion arrière de reconstruire le spectrogramme Mel de la parole propre. Le processus arrière exploite des scores estimés par le réseau de décodage, et ce processus itératif transforme la sortie μ du réseau de conditionnement en un spectrogramme Mel de la parole propre x_0.

L'apprentissage du modèle GradSE proposée utilise deux critères de coût : l'erreur quadratique moyenne (MSE – Mean Square Error) et un critère de diffusion. La coût MSE est appliqué à la sortie du réseau de conditionnement par rapport au spectrogramme Mel de la parole propre. Le bruit résiduel est calculé par échantillonnage à partir d'une distribution de bruit, et le coût MSE est utilisé pour assurer la stabilité de l'apprentissage. Le critère de diffusion est basée sur la divergence de Fisher et est utilisé pour minimiser la divergence entre le gradient de la densité logarithmique du bruit et le gradient prédit par le décodeur U-net basé sur le DPM. Le critère de diffusion permet une meilleure estimation des trajectoires inverses de celles du processus de diffusion avant.

A.4.2 Diff-Estimator

Nous avons proposé un autre modèle, Diff-Estimator, pour estimer les masques temps-fréquence ou les filtres de Wiener multicanal de rang 1 pour le rehaussement de la parole. Le modèle Diff-Estimator consiste en un décodeur de diffusion et ne nécessite pas de réseau de conditionnement. Deux modifications ont été apportées au décodeur basé sur le DPM afin d'améliorer la précision et la robustesse du décodeur pour le traitement des signaux multicanal bruités. Diff-Estimator effectue un processus de diffusion vers l'avant en déconstruisant le masque oracle en une distribution de bruit terminale, qui est paramétrée à l'aide d'un spectrogramme multicanal de bruit. Lors de l'inférence, un spectrogramme multicanal de parole bruitée est donné en entrée à l'ECAPA-TDNN, qui effectue des étapes de diffusion inverse pour calculer le masque ou le filtre MWF de rang 1.

A.4.3 Diff-TasNet

Diff-TasNet est une approche de rehaussement de la parole dans le domaine temporel basée sur le DPM pour la vérification du locuteur. Diff-TasNet remplace le modèle de

diffusion basé sur le réseau U-net par une architecture de réseau neuronal basée sur Conv-TasNet, qui utilise un réseau d'encodage linéaire et un réseau convolutif temporel pour estimer les masques. Diff-TasNet estime les masques dans le domaine temporel et permet l'amélioration de la parole multicanal dans le domaine temporel. L'approche effectue directement l'amélioration de la parole en utilisant le signal brut multicanal dans le domaine temporel comme entrée. Pour garantir la stabilité pendant l'apprentissage, une approche d'apprentissage en deux étapes est utilisée, où le modèle Diff-TasNet est pré-entraîné comme un réseau autoencodeur génératif, puis des signaux vocaux bruités multicanal sont utilisés avec des signaux vocaux propres pour la deuxième partie de l'apprentissage. L'approche génère des trajectoires inverses du processus de diffusion avant pour transformer les signaux bruités multicanal en des signaux vocaux propres.

A.4.4 Diff-Filter

Il a été montré dans la section 4 que le le filtre MWF de rang 1 est une meilleure technique de rehaussement de la parole en comparaison des approches GEV ou de beamforming. Le modèle Diff-Filter proposé est un DPM basé sur les scores qui utilise une architecture Conv-TasNet pour le processus de diffusion.

Diff-Filter comprend un réseau de diffusion et un réseau de conditionnement qui estime les signaux de parole et de bruit à partir des signaux bruités multicanal, qui sont concaténés avec les signaux bruités et le signal de parole propre MWF de rang 1 en tant qu'entrée du réseau de décodeur de diffusion. Le modèle est entraîné à l'aide d'un processus en deux étapes et de deux fonctions de coût sont utilisés, un critère de diffusion et un coût SI-SDR, afin de garantir que le modèle intègre des informations intrinsèques sur les estimations de la parole propre et du bruit. Dans la phase d'inférence, les signaux bruités multicanal et les estimations des signaux de parole et de bruit sont fournis au décodeur de diffusion pour estimer les trajectoires inverses de la diffusion avant.

A.4.5 Vérification du locuteur

Le modèle ECAPA-TDNN a été utilisé pour la vérification du locuteur. Le modèle a été appris sur les ensembles de données VoxCeleb1 et VoxCeleb2 en utilisant l'optimiseur Adam. Les coefficients des spectrogrammes Mel de dimension 40 ont été utilisés comme entrée du modèle, et le critère de vérification du locuteur repose sur le calcul du cosinus entre les embeddings extraits.

A.4.6 Résultats

Performances sur les données RoboVoices

Le tableau A.7 présente les résultats de différentes approches de prétraitement basées sur les modèles probabilistes de diffusion sur l'ensemble de données RoboVoices eval1. Les résultats sont fournis pour 3 niveaux de SNR. Nous avons entraîné les modèles sur les données de RoboVoices. Les trois premières lignes du tableau indiquent les performances pour la parole propre, la parole propre réverbérée et la parole bruitée sans rehaussement. Ces signaux sont utilisés comme signaux de référence pour comparer les performances des approches avec rehaussement de la parole. Les autres lignes du tableau montrent les performances obtenues avec les approches de rehaussement de la parole proposées, basées sur les DPMs.

Table A.7: SIR (dB), SDR (dB) et % EER sur l'ensemble de données RoboVoices eval1 pour différents SNR (dB). Comme il n'est pas possible de calculer le SIR et le SDR sur les spectrogrammes Mel, ces mesures ne sont pas disponibles pour GradSE, qui estime les spectrogrammes Mel. L'intervalle de confiance moyen est de 0,1.

	5			**10**			**20**		
Pre-proces./SNR	**SIR**	**SDR**	**EER**	**SIR**	**SDR**	**EER**	**SIR**	**SDR**	**EER**
Dry clean speech	—	—	5.6	—	—	5.6	—	—	5.6
Reverb. clean speech	31.5	6.7	7.5	31.5	6.7	7.5	31.5	6.7	7.5
Unprocessed (Noisy)	12.9	0.8	11.2	15.1	2.0	9.2	17.4	5.0	7.8
GradSE	—	—	10.2	—	—	8.5	—	—	7.2
Diff-Estimator	21.1	5.2	10.4	20.8	5.9	8.9	26.3	6.2	7.5
ConvTasnet	21.6	5.3	10.2	23.2	6.1	8.6	27.5	6.6	7.1
Diff-TasNet	22.0	5.4	9.9	23.5	6.4	8.4	27.6	6.7	7.0
Diff-Filter	**22.4**	**5.6**	**9.7**	**23.7**	**6.6**	**7.8**	**27.7**	**6.8**	**6.5**

Les valeurs SIR et SDR pour la parole bruitée (sans rehaussement) sont faibles, ce qui indique des interférences et des distorsions importantes. L'EER est élevé pour les SNR 5dB et 10dB. Les approches de rehaussement de la parole proposées, basées sur la diffusion, améliorent les performances par rapport à la parole bruitée (sans rehaussement), comme le montrent les valeurs SIR et SDR plus élevées et les valeurs EER plus faibles. Diff-Filter conduit aux meilleures performances pour toutes les conditions SNR.

Performances sur l'ensemble de données SV multicanal

Table A.8: % EER sur l'ensemble de données MultiSV Eval. La sortie de GradSE est un spectrogramme Mel. L'intervalle de confiance moyen est de 0,1.

	SE	**SV**	**MRE**	**MRE-Hard**
	Unprocessed	ECAPA-TDNN	5.84	10.27
	Oracle Rank-1 MWF	ECAPA-TDNN	1.64	3.12
	GradSE	ECAPA-TDNN	3.77	4.52
Mask	Diff-Estimator	ECAPA-TDNN	3.63	4.57
Mask	Diff-TasNet	ECAPA-TDNN	**3.58**	**4.51**
Mask	Diff-Filter	ECAPA-TDNN	3.65	4.56
Filter	Diff-Estimator	ECAPA-TDNN	3.61	4.55
Filter	Diff-TasNet	ECAPA-TDNN	3.57	4.51
Filter	Diff-Filter	ECAPA-TDNN	**3.53**	**4.36**

Le tableau A.8 présente les résultats des expériences menées sur l'ensemble de données MultiSV dans deux conditions : MRE et MRE Hard. Le tableau montre les valeurs EER moyennes sur tous les essais pour chaque approche. La troisième section du tableau montre les résultats obtenus avec l'estimation de masques temps-fréquence, tandis que la quatrième section montre les résultats obtenus avec l'estimation def iltres multicanal.

La condition MRE-hard, qui présente des niveaux élevés de bruit et de réverbération, conduit à des performances de vérification du locuteur bien plus mauvaises que celles

obtenues pour la condition MRE. Les résultats obtenusmontrent que toutes les approches de rehaussement basées sur la diffusion améliorent les performances de vérification du locuteur, et ce pour les deux conditions de bruit. Les approches basées sur les masques ont amélioré la performance du SV, Diff-TasNet obtenant la meilleure performance parmi les approches basées sur les masques. Cependant, globalement, c'est l'approche Diff-Filter, avec estimation d'un filtre, qui obtient les meilleures performance avec un EER de 3,53% dans la condition MRE et 4,36% dans la condition MRE-hard,, soit une réduction de 6% en absolu du EER pour cette condition difficile.

Validation sur le jeu VOiCES de données publiques

Table A.9: % EER dans différentes conditions de bruit sur l'ensemble de données VOiCES Eval. L'intervalle de confiance moyen est de 0,2.

	Noise conditions			
Testing environment	**Clean**	**Babble**	**TV**	**Music**
Unprocessed	4.1	8.8	7.8	7.9
Oracle Rank-1 MWF	2.8	3.9	4.1	4.2
GradSE	3.9	6.7	6.2	6.6
Diff-Estimator	3.9	6.6	6.6	6.8
Diff-TasNet	3.9	6.3	6.1	6.2
Diff-Filter	**3.8**	**6.2**	**6.0**	**6.0**

Le tableau A.9 compare les performances des modèles basés sur la diffusion pour différentes conditions de bruit de l'ensemble de données VOiCES. La première section du tableau montre les valeurs EER pour le signal non traité et pour un pré-traitement avec l'approche Oracle Rank-1 MWF. La deuxième section du tableau montre les performances des modèles proposés, basés sur la diffusion.

Les résultats indiquent que les modèles basés sur la diffusion améliorent les performances de vérification du locuteur pour toutes les conditions de bruit, l'amélioration la plus importante étant observée pour le bruit babble. Diff-Filter a obtenu la meilleure performance dans toutes les conditions de bruit. Les résultats suggèrent que la performance des approches de rehaussement de la parole est dépendante du type de bruit. Bien qu'ils n'aient pas été entraînés sur l'ensemble de données VOiCES, les modèles basés sur la diffusion ont amélioré notablement les performances de vérification du locuteur.

A.5 Optimisation conjointe des modules de rehaussement de la parole et de vérification du locuteur

L'entraînement indépendant des modèles de rehaussement de parole et de reconnaissance du locuteur peut conduire à des problèmes d'inadéquation ; et le rehaussement de parole peut éventuellement filtrer des caractéristiques utiles pour la reconnaissance du locuteur. L'optimisation conjointe des modules de rehaussement de parole et de vérification du locuteur devrait résoudre ce problème et améliorer les performances de vérification du locuteur en filtrant les bruit et en conservant les informations utiles pour la tâche de vérification du locuteur. Plusieurs études proposent une optimisation conjointe, notamment

la formation de voies MVDR soutenue par un DNN et l'apprentissage WPE, l'extraction d'embedding de locuteur, le masque et le mécanisme d'attention. Toutefois, la plupart de ces études se concentrent sur les données monocanal et sont adaptées au multicanal par le biais d'un traitement supplémentaire.

Les principaux avantages de l'optimisation conjointe sont l'amélioration des performances, la simplification de la conception du système, l'augmentation de la robustesse, l'amélioration de la représentation des caractéristiques et la réduction de la complexité de calcul. Cette section propose deux approches pour l'optimisation conjointe : l'une basée sur un réseau neuronal pour la formation de voies, et l'autre sur les techniques de rehaussement de parole multicanal basées sur les DPMs. Un critère de distillation des connaissance (knowledge distillation loss) est également introduite pour guider l'apprentissage du réseau afin qu'il produise des activations similaires pour les signaux réhaussés et pour les signaux vocaux propres.

A.5.1 Optimisation conjointe

Nous utilisons trois fonctions de coût pour ces expérimentations : l'une associée à SI-SDR, une autre à l'entropie croisée et la dernière à la distillation des connaissances. Le critère SI-SDR, qui est invariant par rapport à l'échelle du signal, mesure l'efficacité des algorithmes de traitement de la parole. Le critère d'entropie croisée est traditionnellement utilisé pour apprendre le module de vérification du locuteur. Enfin, le critère associé à la distillation de connaissances est une technique de compression de modèles qui, typiquement, sert pour extraire les connaissances d'un grand modèle neuronal pré-entraîné et les transfèrer à un petit modèle neuronal, en préservant la similarité des connaissances extraites. Ici, le critère de similarité est appliqué entre les embeddings obtenus après réhaussement de parole, et ceux obtenus directement sur la parole propre.

L'architecture proposée combine les systèmes de rehaussement de parole et de vérification du locuteur et est optimisée comme un modèle unique. Le réseau ECAPA-TDNN est utilisé pour générer des embeddings de locuteurs, qui sont ensuite utilisés pour la classification. Le processus d'optimisation conjointe implique la rétropropagation des gradients d'erreur à travers les réseaux de vérification du locuteur et de rehaussement de la parole en un seul passage. Les systèmes de rehaussement de la parole pré-entraînés, basés sur les DPMs sont utilisés pour l'optimisation conjointe. Les détails du processus d'optimisation conjointe pour les différents systèmes d'amélioration de la parole sont décrits dans les sections suivantes.

A.5.2 FasNet

FaSNet est appris à l'aide du rapport signal-à distorsion invariant à l'échelle (SI-SDR) et de l'entropie croisée. Le signal vocal propre estimé par FaSNet est transmis au réseau ECAPA-TDNN pour apprendre la représentation du locuteur (embedding). Dans la phase d'apprentissage, les deux critères de coût sont pondérés, avec une pondération de 0,1 pour le critère SI-SDR et une pondération de 0,9 pour l'entropie croisée. L'optimiseur Adam est utilisé pour apprendre le réseau global pendant 50 époques.

A.5.3 GradSE

L'optimisation conjointe utilise les critères d'entropie croisée et de distillation des connaissances. L'entrée de GradSE est un spectrogramme Mel multicanal bruité, et GradSE

utilise un processus de diffusion inverse pour reconstruire le spectrogramme Mel propre. Le spectrogramme Mel amélioré est ensuite transmis au réseau ECAPA-TDNN et les gradients d'erreur sont rétropropagés à travers GradSE et ECAPA-TDNN dans la phase d'apprentissage. L'optimiseur Adam, le planificateur de taux d'apprentissage cyclique avec des minibatch de taille 64 et une marge de 0,4 pour la marge angulaire softmax ont été utilisés pour apprendre le réseau pendant 20 000 itérations. Au cours du processus d'optimisation conjointe, une politique triangulaire2 a été utilisée pour le planificateur de taux d'apprentissage cyclique, ainsi qu'un facteur de 30 pour la fonction softmax. Le rehaussement de la parole GradSE repose sur 20 étapes de diffusion inverse pour reconstruire le spectrogramme Mel propre.

A.5.4 Diff-Estimator

Diff-Estimator prend en entrée un spectrogramme de parole bruitée multicanal et, après 20 étapes de diffusion inverse, le réseau décodeur basé sur ECAPA-TDNN un masque estimé pour la parole propre. Les coefficients du spectre Mel sont calculés et transmis au système de vérification du locuteur, et les deux systèmes sont entraînés conjointement à l'aide du critère d'entropie croisée. La phase d'apprentissage conjoint implique 500 itérations avec une taille de minibatch de 16 et un taux d'apprentissage de 0,0001 à l'aide de l'optimiseur Adam.

A.5.5 Diff-Filter

Ce modèle de rehaussement de la parole multicanal basé sur Diff-TasNet utilise convtasNet comme réseau de décodage pour mettre en œuvre un processus de diffusion afin d'éliminer le bruit de manière itérative à partir d'un signal bruité multicanal. Le modèle effectue 20 étapes de diffusion inverse pour obtenir une estimation du signal vocal propre. Ensuite, les coefficients du spectrogramme Mel sont extraits à l'aide de torchaudio et fournies comme entrée au système de vérification du locuteur basé sur ECAPA-TDNN. Les deux systèmes sont entraînés conjointement pendant 300 itérations avec une taille de minibatch de 4 et un taux d'apprentissage de 1e-4 à l'aide de l'optimiseur adam, et la le critère d'entropie croisée.

A.5.6 Résultats

L'efficacité et la robustesse au bruit des modèles optimisés conjointement ont été évaluées expérimentalement sur des ensembles de données simulées et sur des données enregistrées. L'évaluation a été réalisée à l'aide des mesures SIR et SDR pour les performances de rehaussement de la parole et la mesure EER pour les performances de vérification du locuteur. Les résultats ont été calculés sur les ensembles de données RoboVoices eval1, MultiSV et VOiCES. Le évaluations SIR et SDR ne sont pas disponibles pour l'approche GradSE qui produit un spectrogramme Mel débruité. Les performances des approches avec optimisation séparée du rehaussement de parole et de la vérification du locuteur, sont également présentées à des fins de comparaison.

Impact du rapport signal à bruit

Le tableau A.10 présente les résultats d'une étude sur l'ensemble de données RoboVoices à différents niveaux de SNR. Les résultats montrent que le bruit a un impact signifi-

Table A.10: SIR (dB), SDR (dB), et % EER sur l'ensemble de données RoboVoices eval1 pour différents SNR (dB). Comme GradSE fournit un spectrogramme Mel de parole propre, il est impossible de calculer le SIR et le SDR pour cette approche. L'intervalle de confiance moyen est de 0,1.

		5			**10**			**20**		
	Pre-pro./SNR	**SIR**	**SDR**	**EER**	**SIR**	**SDR**	**EER**	**SIR**	**SDR**	**EER**
	Dry clean speech	—	—	5.6	—	—	5.6	—	—	5.6
	Reverb. clean speech	31.5	6.7	7.5	31.5	6.7	7.5	31.5	6.7	7.5
	Oracle Rank-1 MWF	22.7	6.0	5.9	24.7	7.2	5.9	28.2	7.5	5.9
	Unprocessed (noisy)	12.9	0.8	11.2	15.1	2.0	9.2	17.4	5.0	7.8
	Conv-TasNet	21.6	5.3	10.2	23.2	6.1	8.6	27.5	6.6	7.1
	Diff-Estimator	21.1	5.2	10.4	20.8	5.9	8.9	26.3	6.2	7.5
	Diff-TasNet	22.0	5.4	9.9	23.5	6.4	8.4	27.6	6.7	7.0
	Diff-Filter	22.4	5.6	9.7	23.7	6.6	7.8	27.7	6.8	6.5
J. optim.	Conv-TasNet	22.4	5.5	9.3	23.8	6.5	7.8	27.8	6.8	7.2
J. optim.	Diff-Estimator	22.3	5.4	9.6	21.9	6.2	7.9	26.7	6.4	7.1
J. optim.	Diff-TasNet	22.7	5.9	8.8	24.0	6.8	7.5	28.0	7.1	6.6
J. optim.	Diff-Filter	**23.0**	**6.0**	**8.1**	**24.1**	**6.9**	**7.5**	**28.1**	**7.3**	**6.4**

catif sur les performances d'amélioration de la parole. L'approche Diff-Filter a montré les meilleures performances parmi toutes les approches d'amélioration de la parole. L'optimisation conjointe de l'amélioration de la parole et de la SV a amélioré les performances pour toutes les mesures d'évaluation dans toutes les conditions de SNR par rapport aux approches individuelles. Le modèle conjointement optimisé proposé est robuste aux scénarios de faible SNR où le bruit diffus et la réverbération sont activement présents. L'optimisation conjointe a le potentiel d'améliorer la robustesse et l'efficacité des systèmes d'amélioration de la parole et de la SV dans des scénarios réels où le bruit et la réverbération sont présents.

L'optimisation conjointe du rehaussement de la parole et de la vérification du locuteur a amélioré les performances de toutes les mesures d'évaluation (SIR, SDR et EER) dans toutes les conditions de SNR. Les modèles optimisés conjointement et basés sur FaSNet ont obtenu les plus mauvaises performances en termes d'EER.

L'utilisation du critère de distillation des connaissances (KD loss) améliore un peu les performances des modèles FaSNet et GradSE. Les modèles basés sur les DPMs optimisés conjointement, en particulier diff-Filter, ont obtenu les meilleures performances dans toutes les conditions de SNR, surpassant toutes les autres approches.

A.5.7 Impact de la perte de distillation de connaissance

La Table A.11 présente les résultats de deux approches d'amélioration de la parole, FaSNet et GradSE, sur le dataset RoboVoices eval1, en termes de SIR, SDR et EER. Les évaluations ont été effectuées à différents niveaux SNR (5 dB, 10 dB et 20 dB). La deuxième section présente les résultats des approches individuelles, tandis que la dernière section présente les résultats de l'optimisation conjointe des approches d'amélioration de la parole et de la SV basée sur ECAPA-TDNN, avec et sans la perte de distillation des connaissances (KD loss).

Table A.11: SIR (dB), SDR (dB), et % EER sur le jeu de données simulé RoboVoices pour différents SNR (dB) avec et sans la perte de distillation de connaissance. Étant donné que GradSE fonctionne avec des caractéristiques de spectrogramme Mel pour l'amélioration, il est impossible de calculer SIR et SDR à l'aide de spectrogrammes Mel. J. optim. dans le tableau fait référence à l'optimisation conjointe. L'intervalle de confiance moyen est de $0,1$.

		5			**10**			**20**		
	Pre-pro./SNR	**SIR**	**SDR**	**EER**	**SIR**	**SDR**	**EER**	**SIR**	**SDR**	**EER**
	Dry clean speech	—	—	5.6	—	—	5.6	—	—	5.6
	Reverb. clean speech	31.5	6.7	7.5	31.5	6.7	7.5	31.5	6.7	7.5
	Oracle Rank-1 MWF	22.7	6.0	5.9	24.7	7.2	5.9	28.2	7.5	5.9
	Unprocessed (noisy)	12.9	0.8	11.2	15.1	2.0	9.2	17.4	5.0	7.8
	FaSNet	17.0	4.8	12.4	19.7	5.4	10.5	24.6	5.8	8.0
	GradSE	—	—	10.2	—	—	8.5	—	—	7.2
J. optim.	FaSNet	21.1	5.0	10.1	22.8	5.9	8.4	26.2	6.3	7.5
J. optim.	GradSE	—	—	9.8	—	—	8.0	—	—	7.1
J. optim.	FaSNet + KD loss	21.8	5.4	9.9	23.0	6.0	8.0	27.3	6.2	7.1
J. optim.	GradSE + KD loss	—	—	9.2	—	—	7.7	—	—	6.8

Les résultats montrent que l'optimisation conjointe de FaSNet et GradSE avec la SV basée sur ECAPA-TDNN fournit des améliorations significatives par rapport aux approches individuelles en termes d'EER. Les valeurs d'EER sont significativement plus faibles que les approches individuelles à tous les niveaux SNR. Cela indique que l'approche d'optimisation conjointe est efficace pour améliorer les performances de la SV.

Les résultats montrent également que l'ajout de la KD loss à l'approche d'optimisation conjointe améliore encore les performances en termes d'EER. Les valeurs d'EER sont encore plus faibles que l'approche d'optimisation conjointe sans KD loss à tous les niveaux SNR. Cela indique que la KD loss est efficace pour supprimer le bruit et la réverbération pour la SV.

Impact de la perte de distillation de connaissance

Le tableau A.12 présente l'EER moyen pour différentes conditions d'enrôlement sur le dataset MultiSV en utilisant des systèmes d'amélioration de la parole et de vérification de locuteur (SV). Le système SV ReSNet est implémenté pour comparer les performances. Les conditions CE et MRE représentent respectivement un enrôlement clair réverbéré et un enrôlement multi-canal.

L'oracle Rank-1 MWF atteint un EER de 2,7% sur un enrôlement clair, contre 4,0% pour un signal non traité, 1,6% sur MRE contre un EER non traité de 5,8%, et 3,1% sur MRE-hard contre un EER non traité de 10,2%. On observe une amélioration de l'EER grâce à la réverbération présente dans l'enrôlement clair. Les modèles optimisés conjointement ont considérablement amélioré les performances pour toutes les conditions d'enrôlement. Le modèle Diff-Filter a obtenu les meilleures performances dans toutes les conditions d'enrôlement. Dans le cas de MRE-hard, qui présente des sources de bruit de fond difficiles, Diff-Filter a amélioré les performances de vérification de locuteur de près de 6% par rapport au signal non traité. Le modèle Diff-Filter optimisé conjointement est

Table A.12: % EER on the MultSV Eval dataset. Mask and filter in the table refer to the speech enhancement output, which is further given as input to the SV. The output of GradSE is the Mel spectrogram. Joint optim. in the table refers to joint optimization of speech enhancement and SV module. The confidence interval is 0.1.

	SE	**SV**	**CE**	**MRE**	**MRE-Hard**
	Unprocessed	ECAPA-TDNN	4.0	5.8	10.2
	Oracle Rank-1 MWF	ECAPA-TDNN	2.7	1.6	3.1
	Mask	ReSNet	4.2	3.9	5.3
	Conv-TasNet	ReSNet	4.2	3.7	4.6
	Conv-TasNet	ECAPA-TDNN	3.8	3.6	4.5
	Diff-Estimator	ECAPA-TDNN	3.8	3.6	4.5
	GradSE	ECAPA-TDNN	3.8	3.7	4.5
	Diff-TasNet	ECAPA-TDNN	3.6	3.5	4.5
	Diff-Filter	ECAPA-TDNN	3.6	3.5	4.3
Joint optim.	Conv-TasNet	ECAPA-TDNN	3.77	3.61	4.49
Joint optim.	Diff-Estimator	ECAPA-TDNN	3.81	3.58	4.51
Joint optim.	GradSE	ECAPA-TDNN	3.71	3.64	4.43
Joint optim.	Diff-TasNet	ECAPA-TDNN	3.57	3.55	4.34
Joint optim.	Diff-Filter	ECAPA-TDNN	**3.51**	**3.24**	**4.26**

Table A.13: % EER on different noise conditions of the VOiCES Eval dataset. The confidence interval is 0.2.

		Noise conditions			
	Testing environment	**Clean**	**Babble**	**TV**	**Music**
	Unprocessed	4.1	8.8	7.8	7.9
	Oracle Pytorch Rank-1 MWF	2.8	3.9	4.1	4.2
	FaSNet	4.4	7.8	7.4	7.9
	FaSNet Rank-1 MWF WPE	4.2	6.9	6.4	6.8
	GradSE	3.9	6.7	6.2	6.6
	Mask Estimator	4.1	7.1	7.2	7.2
	Diff-Estimator	3.9	6.6	6.6	6.8
	Conv-TasNet	3.9	6.8	6.3	6.4
	Diff-TasNet	3.9	6.3	6.1	6.2
	Diff-Filter	3.8	6.2	6.0	6.0
Joint optimization	FaSNet ECAPA-TDNN	4.0	6.6	6.2	6.5
Joint optimization	GradSE ECAPA-TDNN	3.8	6.4	6.0	6.2
Joint optimization	FaSNet + ECAPA-TDNN + KD loss	3.9	6.6	6.1	6.3
Joint optimization	GradSE + ECAPA-TDNN + KD loss	3.8	6.2	5.9	6.1
Joint optimization	MaskEstimator + ECAPA-TDNN	3.9	6.5	6.2	6.4
Joint optimization	Diff-Estimator + ECAPA-TDNN	3.8	6.4	6.2	6.2
Joint optimization	Conv-TasNet + ECAPA-TDNN	3.8	6.4	6.2	6.3
Joint optimization	Diff-TasNet + ECAPA-TDNN	3.7	6.1	5.7	5.9
Joint optimization	Diff-Filter + ECAPA-TDNN	**3.7**	**5.9**	**5.6**	**5.7**

le meilleur modèle dans toutes les conditions d'enrôlement.

Validation sur le jeu VOiCES de données publiques

Le tableau A.13 montre les résultats des systèmes optimisés conjointement avec et sans perte de distillation des connaissances sur le dataset VOiCES. Les systèmes basés sur DPM sont supérieurs aux systèmes basés sur FaSNet, avec une amélioration constante dans toutes les conditions de bruit. L'utilisation de la perte de distillation des connaissances améliore également les performances des systèmes.

A.6 Conclusion et perspectives

La thèse couvre différentes techniques pour améliorer la performance de la vérification du locuteur dans des environnements bruyants et réverbérants. Nous présentons une revue de la littérature sur les techniques de traitement du signal et les architectures de réseaux neuronaux profonds pour la vérification du locuteur et le rehaussement de la parole. Nous discutons des ensembles de données et des mesures d'évaluation pour les systèmes de rehaussement de la parole et de vérification du locuteur et nous introduisons l'ensemble de données RoboVoices conçu pour le rehaussement de parole multicanal. Nous présentons ensuite une approche pipeline de référence pour le rehaussement de la parole multicanal, combinant des DNN avec des méthodes de traitement du signal, et nous évaluons son impact sur la performance de la vérification du locuteur. Nous avons proposé différentes techniques de rehaussement de la parole multicanal basées sur les processus stochastiques de diffusion, ainsi qu'une approche d'optimisation conjointe des modules de rehaussement de la parole et de vérification du locuteur. Les expériences menées sur différents ensembles de données montrent que l'optimisation conjointe et les techniques basées sur la diffusion améliorent de manière significative les performances de vérification du locuteur dans des environnements acoustiques difficiles.

Les travaux futurs devraient se concentrer sur l'amélioration des performances du système grâce à des modifications des architectures des réseaux, en utilisant de nouvelles procédures d'apprentissage telles que l'apprentissage auto-supervisé, et en rendant les architectures proposées pour le rehaussement de la parole et la vérification du locuteur adaptées au déploiement sur des dispositifs embarqués. Les techniques de déréverbération basées sur les DNN permettront l'optimisation conjointe du rehaussement de la parole, de la déréverbération et de la vérification du locuteur en tant que réseau unique. L'apprentissage auto-supervisé peut permettre l'utilisation de signaux vocaux bruités multicanal non étiquetés en grande quantité. L'apprentissage exploitant la distillation des connaissances peut être appliquée à d'autres systèmes de rehaussement de la parole proposés, et les techniques d'apprentissage en profondeur peuvent être adaptées au déploiement de modèles dans des conditions réelles. Il est également essentiel de préparer les modèles DNN basés sur PyTorch à un déploiement transparent sur diverses plates-formes embarquées, pour lesquelles la bibliothèque ONNX peut être une solution appropriée.

Bibliography

[Abdalmalak and Gallardo-Antolín, 2018] Abdalmalak, K. A. and Gallardo-Antolín, A. (2018). Enhancement of a text-independent speaker verification system by using feature combination and parallel structure classifiers. *Neural Computing and Applications*, 29(3):637–651.

[Ajili et al., 2016] Ajili, M., Bonastre, J., Kahn, J., Rossato, S., and Bernard, G. (2016). Fabiole, a speech database for forensic speaker comparison. In *International Conference on Language Resources and Evaluation (LREC)*.

[Al-Qaderi et al., 2021] Al-Qaderi, M., Lahamer, E., and Rad, A. (2021). A two-level speaker identification system via fusion of heterogeneous classifiers and complementary feature cooperation. *Sensors*, 21(15):5097.

[Alamdari et al., 2021] Alamdari, N., Azarang, A., and Kehtarnavaz, N. (2021). Improving deep speech denoising by noisy2noisy signal mapping. *Applied Acoustics*.

[Anderson, 1982] Anderson, B. (1982). Reverse-time diffusion equation models. *Stochastic Processes and their Applications*.

[Bahl, 1993] Bahl, L. R., B. P. F. d. S. P. V. . M. R. L. (1993). Speaker verification using adapted hmm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pages 746–758.

[Bai and Zhang, 2020] Bai, Z. and Zhang, X.-L. (2020). Speaker recognition based on deep learning: An overview. *Neural networks: the official journal of the International Neural Network Society*, 140:65–99.

[Balan and Rosca, 2002] Balan, R. and Rosca, J. (2002). Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase. In *Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002*, pages 209–213.

[Bando et al., 2017] Bando, Y., Mimura, M., Itoyama, K., Yoshii, K., and Kawahara, T. (2017). *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 716–720.

[Bando et al., 2020] Bando, Y., Sekiguchi, K., and Yoshii, K. (2020). Adaptive neural speech enhancement with a denoising variational autoencoder. In *Interspeech*.

[Barker et al., 2016] Barker, J., Marxer, R., Vincent, E., and Watanabe, S. (2016). The third 'chime' speech separation and recognition challenge: Analysis and outcomes. *Computer Speech Language*, 46.

[Barker et al., 2013] Barker, J., Vincent, E., Ma, N., Christensen, H., and Green, P. (2013). The PASCAL CHiME speech separation and recognition challenge. *Computer Speech and Language*, 27(3):621–633.

[Benesty and Huang, 2013] Benesty, J. and Huang, Y. (2013). *Adaptive signal processing: applications to real-world problems*. Springer Science & Business Media.

[Bian et al., 2019] Bian, T., Chen, F., and Xu, L. (2019). Self-attention based speaker recognition using cluster-range loss. *Neurocomputing*, 368:59–68.

[Bisani and Ney, 2004] Bisani, M. and Ney, H. (2004). Bootstrap estimates for confidence intervals in asr performance evaluation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages I–409.

[Boll, 1979] Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120.

[Bourlard, 1998] Bourlard, H., . B. N. (1998). Joint factor analysis vs. mllr in speaker verification. pages 1033–1036.

[Cai et al., 2018] Cai, W., Chen, J., and Li, M. (2018). Exploring the encoding layer and loss function in end-to-end speaker and language recognition system. *arXiv preprint arXiv:1804.05160*.

[Campbell, 1997] Campbell, J. (1997). Speaker recognition: a tutorial. *Proceedings of IEEE*, 85:1437–1462.

[Campbell et al., 2006] Campbell, W. M., Sturim, D. E., and Reynolds, D. A. (2006). Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, 13:308–311.

[Capon, 1969] Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418.

[Cappe, 1994] Cappe, O. (1994). Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, 2(2):345–349.

[Chang and Wang, 2017] Chang, J. and Wang, D. (2017). Robust speaker recognition based on dnn/i-vectors and speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5415–5419.

[Chen et al., 2007] Chen, H., Ser, W., and Yu, Z. L. (2007). Optimal design of nearfield wideband beamformers robust against errors in microphone array characteristics. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 54:1950–1959.

[Chen and Salman, 2011] Chen, K. and Salman, A. (2011). Learning speaker-specific characteristics with a deep neural architecture. *IEEE Transactions on Neural Networks*, 22(11):1744–1756.

[Chen et al., 2020] Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., and Chan, W. (2020). Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*.

[Cho et al., 2014] Cho, K., Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation.

[Chung et al., 2018] Chung, J. S., Nagrani, A., and Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. *Interspeech*.

[Cornelis et al., 2011] Cornelis, B., Moonen, M., and Wouters, J. (2011). Performance analysis of multichannel wiener filter-based noise reduction in hearing aids under second order statistics estimation errors. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1368–1381.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20:273–297.

[Davis and Mermelstein, 1980] Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

[Defferrard et al., 2016] Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. (2016). Fma: A dataset for music analysis. In *International Society for Music Information Retrieval Conference(ISMIR)*.

[Dehak et al., 2011] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

[Delebecque et al., 2022] Delebecque, L., Serizel, R., and Furnon, N. (2022). Towards an efficient computation of masks for multichannel speech enhancement. In *European Signal Processing Conference (EUSIPCO)*.

[Desplanques et al., 2020] Desplanques, B., Thienpondt, J., and Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech*.

[Doclo et al., 2010] Doclo, S., Gannot, S., Moonen, M., Spriet, A., Haykin, S., and Liu, K. R. (2010). Acoustic beamforming for hearing aid applications. *Handbook on array processing and sensor networks*, pages 269–302.

[Doclo et al., 2006] Doclo, S., Klasen, T. J., Van den Bogaert, T., Wouters, J., and Moonen, M. (2006). Theoretical analysis of binaural cue preservation using multi-channel wiener filtering and interaural transfer functions. In *International Workshop Acoustic Echo Noise Control (IWAENC)*, pages 1–4.

[Doclo and Moonen, 2002a] Doclo, S. and Moonen, M. (2002a). Gsvd-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Transactions on Signal Processing*, 50(9):2230–2244.

[Doclo and Moonen, 2002b] Doclo, S. and Moonen, M. (2002b). Gsvd-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Transactions on Signal Processing*, 50(9):2230–2244.

[Doddington, 1998] Doddington, G., M. A. P. M. . R. D. (1998). The nist speaker recognition evaluation: Overview, methodology, systems, results, perspective. *Speech communication, Speech and Signal Processing*, pages 3–28.

[Dufera and Shimamura, 2009] Dufera, B. D. and Shimamura, T. (2009). Reverberated speech enhancement using neural networks. In *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 441–444.

[Ephraim and Malah, 1984] Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121.

[Fang et al., 2020] Fang, X., Gao, T., Zou, L., and Ling, Z. (2020). Bidirectional attention for text-dependent speaker verification. *Sensors*, 20(23):6784.

[Févotte et al., 2005] Févotte, C., Gribonval, R., and Vincent, E. (2005). Bss_eval toolbox user guide – revision 2.0.

[Fisher, 1986] Fisher, W. (1986). The darpa speech recognition research database: specifications and status. In *DARPA Workshop on speech recognition*, pages 93–99.

[Frost, 1972] Frost, O. (1972). An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8):926–935.

[Fu et al., 2017] Fu, S.-W., Tsao, Y., Lu, X., and Kawai, H. (2017). Raw waveform-based speech enhancement by fully convolutional networks. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 006–012.

[Furnon et al., 2020] Furnon, N., Serizel, R., Illina, I., and Essid, S. (2020). Dnn-based distributed multichannel mask estimation for speech enhancement in microphone arrays. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4672–4676.

[Garcia-Romero et al., 2019] Garcia-Romero, D., Snyder, D., Sell, G., McCree, A., Povey, D., and Khudanpur, S. (2019). x-vector dnn refinement with full-length recordings for speaker recognition. In *Interspeech*.

[Ghalehjegh and Rose, 2015] Ghalehjegh, S. H. and Rose, R. C. (2015). Deep bottleneck features for i-vector based text-independent speaker verification. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 555–560.

[Giri et al., 2019] Giri, R., Isik, U., and Krishnaswamy, A. (2019). Attention wave-u-net for speech enhancement. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 249–253.

[Goodfellow et al., 2014] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial nets. In *NeurIPS*.

[Gopinath, 2001] Gopinath, M. A., . S. S. (2001). Improving the robustness of hmm-based speaker verification systems. pages 793–796.

[Greig et al., 1989] Greig, D., Porteous, B., and Seheult, A. H. (1989). Exact maximum a posteriori estimation for binary images. *Journal of the royal statistical society series b-methodological*, 51:271–279.

[Griffiths and Jim, 1982] Griffiths, L. and Jim, C. (1982). An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1):27–34.

[Haykin, 2002] Haykin, S. S. (2002). *Adaptive filter theory*. Pearson Education India.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[Heck and Genoud, 2001] Heck, L. P. and Genoud, D. (2001). Integrating speaker and speech recognizers: Automatic identity claim capture for speaker verification. In *Odyssey*, pages 249–254.

[Heigold et al., 2016] Heigold, G., Moreno, I., Bengio, S., and Shazeer, N. (2016). End-to-end text-dependent speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5115–5119.

[Heymann et al., 2015] Heymann, J., Drude, L., Chinaev, A., and Haeb-Umbach, R. (2015). Blstm supported gev beamformer front-end for the 3rd chime challenge. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 444–451.

[Heymann et al., 2016] Heymann, J., Drude, L., and Häb-Umbach, R. (2016). Neural network-based spectral mask estimation for acoustic beamforming. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[Hinton et al., 2015] Hinton, G., Vinyals, O., Dean, J., et al. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

[Hirsch and Pearce, 2000] Hirsch, H.-G. and Pearce, D. (2000). The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Automatic speech recognition: challenges for the new Millenium ISCA tutorial and research workshop (ITRW)*.

[Ho et al., 2020] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Hu and Loizou, 2003] Hu, Y. and Loizou, P. (2003). A perceptually motivated approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 11(5):457–465.

[Hu and Loizou, 2004] Hu, Y. and Loizou, P. (2004). Incorporating a psychoacoustical model in frequency domain speech enhancement. *IEEE Signal Processing Letters*, 11(2):270–273.

[Hu and Loizou, 2007] Hu, Y. and Loizou, P. C. (2007). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1):229–238.

[Huang, 1991] Huang, X. (1991). A study on speaker-adaptive speech recognition. In *Human Language Technology - The Baltic Perspective*.

[Jeong et al., 2021] Jeong, M., Kim, H., Cheon, S. J., Choi, B. J., and Kim, N. S. (2021). Diff-tts: A denoising diffusion model for text-to-speech. *ArXiv*, abs/2104.01409.

[Jin and Yoo, 2010] Jin, M. and Yoo, C. D. (2010). Speaker verification and identification.

[Juang, 1991] Juang, B. H., . R. L. (1991). A comparative study of hmm and ann-based speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pages 590–604.

[Kanagasundaram et al., 2016] Kanagasundaram, A., Dean, D., Sridharan, S., and Fookes, C. (2016). Domain adaptation based speaker recognition on short utterances. *ArXiv*, abs/1610.02831.

[Kenny, 2005] Kenny, P. (2005). Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM*, 14(28-29):2.

[Kenny et al., 2003] Kenny, P., Mihoubi, M., and Dumouchel, P. (2003). New map estimators for speaker recognition. *European Conference on Speech Communication and Technology (Eurospeech)*.

[Kenny et al., 2013] Kenny, P., Stafylakis, T., Ouellet, P., Alam, M. J., and Dumouchel, P. (2013). Plda for speaker verification with utterances of arbitrary duration. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7649–7653.

[Kinnunen and Li, 2010] Kinnunen, T. H. and Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52:12–40.

[Kloeden and Platen, 1977a] Kloeden, P. E. and Platen, E. (1977a). Numerical solution of stochastic differential equations. In *Applications of Mathematics book series*, volume 23.

[Kloeden and Platen, 1977b] Kloeden, P. E. and Platen, E. (1977b). Numerical solution of stochastic differential equations. In *Applications of Mathematics book series*.

[Kolbæk et al., 2019] Kolbæk, M., Tan, Z., Jensen, S. H., and Jensen, J. H. (2019). On loss functions for supervised monaural time-domain speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:825–838.

[Kowalski et al., 2010] Kowalski, M., Vincent, E., and Gribonval, R. (2010). Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1818–1829.

[Kuhn et al., 1998] Kuhn, R., Nguyen, P., Junqua, J.-C., Goldwasser, L., Niedzielski, N., Fincke, S., Field, K. L., and Contolini, M. (1998). Eigenvoices for speaker adaptation. *International Conference on Spoken Language Processing (ICSLP)*.

[Larcher et al., 2013] Larcher, A., Lee, K. A., Ma, B., and Li, H. (2013). Phonetically-constrained plda modeling for text-dependent speaker verification with multiple short utterances. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7673–7677.

[Le Prell and Clavier, 2016] Le Prell, C. and Clavier, O. (2016). Effects of noise on speech recognition: Challenges for communication by service members. *Hearing Research*, 349.

[Le Roux et al., 2019] Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. (2019). Sdr–half-baked or well done? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630.

[Lea et al., 2016] Lea, C. S., Vidal, R., Reiter, A., and Hager, G. (2016). Temporal convolutional networks: A unified approach to action segmentation. *ArXiv*, abs/1608.08242.

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

[Lehtinen et al., 2018] Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., and Aila, T. (2018). Noise2noise: Learning image restoration without clean data. *ArXiv*, abs/1803.04189.

[Li et al., 2017] Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., and Zhu, Z. (2017). Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*.

[Ljung, 1998] Ljung, L. (1998). System identification. In *Signal analysis and prediction*, pages 163–173. Springer.

[Loizou, 2007] Loizou, P. C. (2007). *Speech enhancement: theory and practice*. CRC press.

[Lu et al., 2021] Lu, Y.-J., Tsao, Y., and Watanabe, S. (2021). A study on speech enhancement based on diffusion probabilistic model. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 659–666.

[Lu et al., 2022] Lu, Y.-J., Wang, Z., Watanabe, S., Richard, A., Yu, C., and Tsao, Y. (2022). Conditional diffusion probabilistic model for speech enhancement. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7402–7406.

[Luo et al., 2019] Luo, Y., Ceolini, E., Han, C., Liu, S.-C., and Mesgarani, N. (2019). Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 260–267.

[Luo et al., 2020] Luo, Y., Chen, Z., and Yoshioka, T. (2020). Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 46–50.

[Mandasari et al., 2011] Mandasari, M. I., McLaren, M., and van Leeuwen, D. A. (2011). Evaluation of i-vector speaker recognition systems for forensic application. In *Interspeech*.

[Markel et al., 1977] Markel, J., Oshika, B., and Gray, A. (1977). Long-term feature averaging for speaker recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(4):330–337.

[McAulay and Malpass, 1980] McAulay, R. and Malpass, M. (1980). Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(2):137–145.

[McLaren et al., 2016] McLaren, M., Ferrer, L., Castán Lavilla, D., and Lawson, A. (2016). The speakers in the wild (sitw) speaker recognition database. In *Interspeech*.

[Michelsanti and Tan, 2017] Michelsanti, D. and Tan, Z. (2017). Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. In *Interspeech*.

[Ming et al., 2007] Ming, J., Hazen, T. J., Glass, J. R., and Reynolds, D. A. (2007). Robust speaker recognition in noisy conditions. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1711–1723.

[Minsky, 1961] Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30.

[Mošner et al., 2022] Mošner, L., Plchot, O., Burget, L., and Černockỳ, J. H. (2022). Multisv: Dataset for far-field multi-channel speaker verification.

[Mošner et al., 2018] Mošner, L., Matějka, P., Novotný, O., and Černocký, J. H. (2018). Dereverberation and beamforming in far-field speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5254–5258.

[Muckenhirn et al., 2017] Muckenhirn, H., Magimai-Doss, M., and Marcel, S. (2017). End-to-end convolutional neural network-based voice presentation attack detection. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 335–341.

[Nagrani et al., 2017] Nagrani, A., Chung, J. S., and Zisserman, A. (2017). Voxceleb: A large-scale speaker identification dataset. *Interspeech*.

[Nandwana et al., 2019] Nandwana, M. K., van Hout, J., Richey, C., McLaren, M., Barrios, M., and Lawson, A. (2019). The voices from a distance challenge 2019. In *Interspeech*.

[Naylor and Gaubitch, 2011] Naylor, P. and Gaubitch, N. (2011). *Speech Dereverberation*. Springer.

[Nugraha et al., 2016] Nugraha, A. A., Liutkus, A., and Vincent, E. (2016). Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1652–1664.

[Oglesby, 1995] Oglesby, J. (1995). What's in a number? moving beyond the equal error rate. *Speech Communication*, 17(1):193–208.

[Openshaw and Masan, 1994] Openshaw, J. and Masan, J. (1994). On the limitations of cepstral features in noise. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 49–52.

[Oppenheim and Schafer, 2009] Oppenheim, A. V. and Schafer, R. W. (2009). *Discrete-Time Signal Processing*. Prentice Hall Press, USA, 3rd edition.

[Ortega-Garcia and González-Rodríguez, 1996] Ortega-Garcia, J. and González-Rodríguez, J. (1996). Overview of speech enhancement techniques for automatic speaker recognition. *Fourth International Conference on Spoken Language Processing (ICSLP)*.

[Panayotov et al., 2015] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

[Pandey and Wang, 2019] Pandey, A. and Wang, D. (2019). Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6875–6879.

[Pandey and Wang, 2020a] Pandey, A. and Wang, D. (2020a). Dense cnn with self-attention for time-domain speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1270–1279.

[Pandey and Wang, 2020b] Pandey, A. and Wang, D. (2020b). Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6629–6633.

[Pandey and Wang, 2020c] Pandey, A. and Wang, D. (2020c). Dual-path self-attention rnn for real-time speech enhancement. *ArXiv*, abs/2010.12713.

[Pariente et al., 2020] Pariente, M., Cornell, S., Cosentino, J., Sivasankaran, S., Tzinis, E., Heitkaemper, J., Olvera, M., Stöter, F.-R., Hu, M., Martın-Doñas, J. M., Ditter, D., Frank, A., Deleforge, A., and Vincent, E. (2020). Asteroid: the PyTorch-based audio source separation toolkit for researchers. In *Interspeech*.

[Park and Lee, 2016] Park, S. R. and Lee, J. (2016). A fully convolutional neural network for speech enhancement. *ArXiv*, abs/1609.07132.

[Pascual et al., 2017] Pascual, S., Bonafonte, A., and Serrà, J. (2017). Segan: Speech enhancement generative adversarial network. In *Interspeech*.

[Peddinti, 2015] Peddinti, V., C. G. M. V. K. T. P. D. K. S. (2015). Jhu aspire system: robust lvcsr with tdnns, i-vector adaptation, and rnn-lms. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

[Peddinti et al., 2015] Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Interspeech*.

[Perotin et al., 2018] Perotin, L., Serizel, R., Vincent, E., and Guérin, A. (2018). Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 36–40.

[Plchot et al., 2016] Plchot, O., Burget, L., Aronowitz, H., and Matejka, P. (2016). Audio enhancing with dnn autoencoder for speaker recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5090–5094.

[Poddar et al., 2018] Poddar, A., Sahidullah, M., and Saha, G. (2018). Speaker verification with short utterances: a review of challenges, trends and opportunities. *IET Biometrics*, 7(2):91–101.

[Popov et al., 2021] Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., and Kudinov, M. (2021). Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning (ICML)*.

[Povey et al., 2011] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

[Prince and Elder, 2007] Prince, S. J. and Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In *IEEE International Conference on Computer Vision*, pages 1–8.

[Qian et al., 2017] Qian, K., Zhang, Y., Chang, S., Yang, X., Florêncio, D. A. F., and Hasegawa-Johnson, M. A. (2017). Speech enhancement using bayesian wavenet. In *Interspeech*.

[Qin et al., 2020] Qin, X., Li, M., Bu, H., Rao, W., Das, R. K., Narayanan, S. S., and Li, H. (2020). The interspeech 2020 far-field speaker verification challenge. In *Interspeech*.

[Rabiner, 1989] Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[Rabiner and Juang, 1993a] Rabiner, L. and Juang, B. (1993a). *Fundamentals of Speech Recognition*. Prentice-Hall Signal Processing Series.

[Rabiner and Juang, 1993b] Rabiner, L. and Juang, B.-H. (1993b). *Fundamentals of speech recognition*. Prentice-Hall, Inc.

[rahman Chowdhury et al., 2018] rahman Chowdhury, F. R., Wang, Q., Moreno, I. L., and Wan, L. (2018). Attention-based models for text-dependent speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5359–5363.

[Rethage et al., 2017] Rethage, D., Pons, J., and Serra, X. (2017). A wavenet for speech denoising. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5069–5073.

[Reynolds et al., 2000] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:19–41.

[Reynolds and Rose, 1995] Reynolds, D. A. and Rose, R. C. (1995). Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions Speech Audio Processing*, 3:72–83.

[Ribas et al., 2022] Ribas, D., Miguel, A., Ortega, A., and Lleida, E. (2022). Wiener filter and deep neural networks: A well-balanced pair for speech enhancement. *Applied Sciences*, 12:9000.

[Richey et al., 2018] Richey, C., Artigas, M., Armstrong, Z., Bartels, C. D., Franco, H., Graciarena, M., Lawson, A. D., Nandwana, M. K., Stauffer, A. R., van Hout, J., Gamble, P., Hetherly, J., Stephenson, C., and Ni, K. S. (2018). Voices obscured in complex environmental settings (voices) corpus. *ArXiv*, abs/1804.05053.

[Richter et al., 2020] Richter, J., Carbajal, G., and Gerkmann, T. (2020). Speech enhancement with stochastic temporal convolutional networks. In *Interspeech*.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference On Medical Image Computing Computer Assisted Intervention. (MICCAI)*.

[Rothauser, 1969] Rothauser, E. (1969). Ieee recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246.

[Sadjadi and Hansen, 2010] Sadjadi, S. O. and Hansen, J. H. (2010). Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions. In *Eleventh Annual Conference of the International Speech Communication Association*.

[Scheibler et al., 2018] Scheibler, R., Bezzam, E., and Dokmanić, I. (2018). Pyroomacoustics: A python package for audio room simulation and array processing algorithms. pages 351–355.

[Serizel et al., 2013] Serizel, R., Moonen, M., van Dijk, B., and Wouters, J. (2013). Rank-1 approximation based multichannel wiener filtering algorithms for noise reduction in cochlear implants. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8634–8638.

[Serrà et al., 2022] Serrà, J., Pascual, S., Pons, J., Araz, R. O., and Scaini, D. (2022). Universal speech enhancement with score-based diffusion. *arXiv preprint arXiv:2206.03065*.

[Shahin et al., 2021] Shahin, I., Nassif, A. B., Nemmour, N., Elnagar, A., Alhudhaif, A., and Polat, K. (2021). Novel hybrid dnn approaches for speaker verification in emotional and stressful talking environments. *Neural Computing and Applications*, 33(23):16033–16055.

[Shi et al., 2020] Shi, Y., Huang, Q., and Hain, T. (2020). Robust speaker recognition using speech enhancement and attention model. *Odyssey*.

[Shon, 2015] Shon, S. (2015). Text independent speaker verification using dominant state information of hmm-ubm. *The Journal of the Acoustical Society of Korea*, 34.

[Shon et al., 2019] Shon, S., Tang, H., and Glass, J. (2019). Voiceid loss: Speech enhancement for speaker verification. *arXiv preprint arXiv:1904.03601*.

[Simmer et al., 2001] Simmer, K. U., Bitzer, J., and Marro, C. (2001). Post-filtering techniques. In *Microphone Arrays*.

[Snyder et al., 2015] Snyder, D., Chen, G., and Povey, D. (2015). Musan: A music, speech, and noise corpus. *ArXiv*, abs/1510.08484.

[Snyder et al., 2018] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

[Sohl-Dickstein et al., 2015] Sohl-Dickstein, J. N., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *ArXiv*, abs/1503.03585.

[Song et al., 2020a] Song, Y., Dickstein, J. S., Kingma, D. P., Kumar, A., Ermon, S., and B.Poole (2020a). Score-based generative modelling through stochastic differential equations. *International Conference on Learning Representations (ICLR)*.

[Song et al., 2021a] Song, Y., Durkan, C., Murray, I., and Ermon, S. (2021a). Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428.

[Song et al., 2021b] Song, Y., Durkan, C., Murray, I., and Ermon, S. (2021b). Maximum likelihood training of score-based diffusion models. In *NeurIPS*.

[Song et al., 2020b] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020b). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

[Souden et al., 2010] Souden, M., Benesty, J., and Affes, S. (2010). On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):260–276.

[Srinivasan et al., 2006] Srinivasan, S., Samuelsson, J., and Kleijn, W. (2006). Codebook driven short-term predictor parameter estimation for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):163–176.

[Stevens et al., 1937] Stevens, S. S., Volkmann, J. E., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8:185–190.

[Sztah'o et al., 2019] Sztah'o, D., Szasz'ak, G., and Beke, A. (2019). Deep learning methods in speaker recognition: a review. *ArXiv*, abs/1911.06615.

[Taherian et al., 2019a] Taherian, H., Wang, Z. Q., and Wang, D. (2019a). Deep learning based multi-channel speaker recognition in noisy and reverberant environments. In *Interspeech*.

[Taherian et al., 2019b] Taherian, H., Wang, Z.-Q., and Wang, D. (2019b). Deep learning based multi-channel speaker recognition in noisy and reverberant environments. In *Interspeech*.

[Thiemann et al., 2013] Thiemann, J., Ito, N., and Vincent, E. (2013). The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. *The Journal of the Acoustical Society of America*, 133:3591.

[Tu et al., 2014] Tu, Y., Du, J., Xu, Y., Dai, L., and Lee, C.-H. (2014). Deep neural network based speech separation for robust speech recognition. In *International Conference on Signal Processing (ICSP)*, pages 532–536.

[Tung and Mori, 2019] Tung, F. and Mori, G. (2019). Similarity-preserving knowledge distillation. In *IEEE/CVF International Conference on Computer Vision*, pages 1365–1374.

[Van Trees, 1969] Van Trees, H. L., . K. T. (1969). Detection, estimation, and modulation theory: Part iv - optimum array processing. volume 57(11), pages 1701–1724.

[Van Veen and Buckley, 1988] Van Veen, B. and Buckley, K. (1988). Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24.

[Variani et al., 2014] Variani, E., Lei, X., McDermott, E., Moreno, I. L., and Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4052–4056.

[Veen and Buckley, 1988] Veen, B. D. V. and Buckley, K. (1988). Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5:4–24.

[Vincent et al., 2006] Vincent, E., Gribonval, R., and Fevotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469.

[Vincent et al., 2016] Vincent, E., Watanabe, S., Nugraha, A., Barker, J., and Marxer, R. (2016). An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech Language*, 46.

[Waibel et al., 1989] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339.

[Wan et al., 2018] Wan, L., Wang, Q., Papir, A., and Moreno, I. L. (2018). Generalized end-to-end loss for speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883.

[Wang et al., 2017] Wang, Z., Vincent, E., Serizel, R., and Yan, Y. (2017). Rank-1 constrained multichannel wiener filter for speech recognition in noisy environments. *Computer Speech and Language*, 49:37–51.

[Wang and Wang, 2016] Wang, Z.-Q. and Wang, D. (2016). A joint training framework for robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):796–806.

[Warsitz and Haeb-Umbach, 2007] Warsitz, E. and Haeb-Umbach, R. (2007). Blind acoustic beamforming based on generalized eigenvalue decomposition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1529–1539.

[Welker et al., 2022] Welker, S., Richter, J., and Gerkmann, T. (2022). Speech enhancement with score-based generative models in the complex stft domain. *arXiv preprint arXiv:2203.17004*.

[Weninger et al., 2015] Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Roux, J. L., Hershey, J. R., and Schuller, B. (2015). Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *Latent Variable Analysis and Signal Separation*, pages 91–99.

[Widrow et al., 1967] Widrow, B., Mantey, P., Griffiths, L., and Goode, B. (1967). Adaptive antenna systems. *Proceedings of the IEEE*, 55(12):2143–2159.

[Wiener et al., 1949] Wiener, N., Wiener, N., Mathematician, C., Wiener, N., Wiener, N., and Mathématicien, C. (1949). *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*, volume 113. MIT press Cambridge, MA.

[Woodbury, 1950] Woodbury, M. A. (1950). Inverting modified matrices. *Statistical Research Group*.

[Xiao et al., 2017] Xiao, X., Zhao, S., Jones, D. L., Chng, E. S., and Li, H. (2017). On time-frequency mask estimation for mvdr beamforming with application in robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3246–3250.

[Xu et al., 2014a] Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2014a). An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*, 21(1):65–68.

[Xu et al., 2014b] Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2014b). Global variance equalization for improving deep neural network based speech enhancement. In *IEEE China Summit International Conference on Signal and Information Processing (ChinaSIP)*, pages 71–75.

[Xu et al., 2015] Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2015). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):7–19.

[Yang and Chang,] Yang, J.-Y. and Chang, J.-H. Joint optimization of neural acoustic beamforming and dereverberation with x-vectors for robust speaker verification. In *Interspeech*.

[Yoshioka and Nakatani, 2012] Yoshioka, T. and Nakatani, T. (2012). Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(10):2707–2720.

[Yu et al., 2019] Yu, Y.-Q., Fan, L., and Li, W.-J. (2019). Ensemble additive margin softmax for speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6046–6050.

[Zeinali et al., 2017] Zeinali, H., Sameti, H., and Burget, L. (2017). Hmm-based phrase-independent i-vector extractor for text-dependent speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP:1–1.

[Zhang, 2017] Zhang, C. (2017). Joint training methods for tandem and hybrid speech recognition systems using deep neural networks.

[Zhang et al., 2018] Zhang, C., Koishida, K., and Hansen, J. H. (2018). Text-independent speaker verification based on triplet convolutional neural network embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9):1633–1644.

[Zhao et al., 2019] Zhao, F., Li, H., and Zhang, X. (2019). A robust text-independent speaker verification method based on speech separation and deep speaker. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6101–6105.

[Zhao et al., 2014] Zhao, X., Wang, Y., and Wang, D. (2014). Robust speaker identification in noisy and reverberant conditions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3997–4001.