Time-domain Multi-channel Speech Separation for Overlapping Speech Recognition

Jisi Zhang

Supervisor: Prof. Jon Barker

Department of Computer Science
University of Sheffield

This thesis is submitted for the degree of
Doctor of Philosophy

July, 2022

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this work are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text.

Acknowledgements

It is my great fortune to have had Prof. Jon Barker be my PhD supervisor. Jon is an excellent supervisor who has provided me with numerous insightful comments on my research and helped me a lot to go through the (Covid-19) pandemic period. Without his help, I would not have completed the study. In addition, I must also thank Prof. Steve Renals for his recommendation to help me get the PhD position.

Many thanks to my PhD panel members: Prof. Thomas Hain and Prof. Lucia Specia, for their suggestions and encouragement. I would also like to thank my thesis examiners: Dr Peter Bell and Dr Yoshi Gotoh, for their time.

Special thanks to Dr Feifei Xiong for helping me with my early study at the beginning of my PhD journey and daily life in Sheffield. Thanks to Dr Catalin Zorila and Dr Rama Doddipatla for their supervision during my internship in Toshiba Research Laboratory and the later collaboration. I want to acknowledge the financial support from Toshiba Europe Limited.

Thanks to my special friend: Dr Shucong Zhang, for all the exciting discussions during the internship in Cambridge. I would also like to thank all my friends in the Department of Computer Science at the University of Sheffield. In particular, thanks to Jack, Gerardo, Zehai, Mingjie, Yanpei, Haiyang, Yilin, Zhengjun and Wanli for the happy moments spent together.

Special thanks to Yidie Cheng for staying with me throughout the long PhD journey. Thanks to her for encouraging me every day through video chatting and ensuring that I finished the thesis.

Finally, thanks to my mother for her endless love.

Abstract

Despite the recent progress of automatic speech recognition (ASR) driven by deep learning, conversational speech recognition using distant microphones is still challenging. In natural environments, an utterance recorded by distant microphones is corrupted by noise and reverberation, and overlapped by competing speakers, which degrade the speech recognition performance.

Speech separation techniques aim to recover individual sources from a noisy mixture, and have been shown beneficial to robust ASR. Deep-learning based separation approaches using a single microphone have moved towards directly processing time-domain signals and outperformed time-frequency domain approaches. When multiple microphones are available, spatial information has been demonstrated to be beneficial for separation. This thesis investigates deep-learning based approaches for time-domain separation using multiple microphones. The designed system is further applied to overlap speech recognition in noisy environments. Three major contributions are summarised as follows.

Firstly, a fully-convolutional multi-channel time-domain separation network is developed. The system uses a neural network to automatically learn spatial features from multiple recordings. Different network architectures and multi-stage separation are also considered for the system design. Experiments show that the proposed system achieves better separation and recognition performance over a conventional time-frequency domain approach.

Next, the time-domain separation system is extended to a speaker extraction system, which employs speaker identity information. A two-stage speaker conditioning mechanism is proposed to efficiently inform the speaker information to the extraction system. The proposed extraction system can simultaneously output multiple corresponding sources from a noisy mixture and further improve the recognition performance over the blind separation approach.

The third contribution studies unsupervised and semi-supervised learning approaches to establish a separation system in situations where only a limited amount of clean data is accessible. An existing unsupervised training strategy that trains a separation system to predict mixtures is improved by exploiting teacher-student learning approaches in this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

The field of automatic speech recognition (ASR) is concerned with the development of systems that can take audio speech recordings and translate them into text transcriptions. In recent years, this technology has become highly effective, especially when using close-talking microphones that are positioned very close to a speaker's mouth. Recent rapid progress can be attributed to multiple factors including the increasing availability of large-scale audio data [Panayotov et al., 2015], advances in deep learning technologies [Hinton et al., 2012], and increasingly powerful computational resources. The progress is such that, in close-talking microphone scenarios, machine recognition performances have now reached near human-level performances for both prepared speech [Povey et al., 2018] and conversational telephone speech [Saon et al., 2017].

However, *distant speech recognition*, which uses audio signals recorded by microphones located at a distance to speakers, remains a largely unsolved problem. Notably, there remains a considerable performance gap between close-talking microphone and distant microphone ASR performance [Renals and Swietojanski, 2017]. Addressing this problem is important because effective distant microphone ASR would provide many new application opportunities. For example, compared to ASR using close-talking microphones, distant speech recognition allows a more natural human computer interface for applications such as voice-control of smart home devices, interactions between humans and robots, automatic transcription of meetings and so on.

The difficulty in distant speech recognition mainly comes from interfering factors in everyday environments. These factors include ambient noise, reverberation, and overlapping speech from simultaneous speakers. Much progress has been made to improve the robustness of speech recognition systems to noise [Barker et al., 2017] and reverberation [Kinoshita et al., 2016]. However, there is one factor that remains a particular challenge even to the current state-of-the-art, namely, overlapping speech [Barker et al., 2018]. To gain an intuition for how overlapping speech affects an ASR system, we can consider how overlapping speech affects human perception. From the perspective of human speech perception, interfering speech can produce two different types of masking effect on a target utterance [Brungart, 2005]. One is that the interference causes an 'energetic' masking when energy from the masker and speech co-occur at the same time at the same frequencies. This results in information loss in the auditory periphery. This phenomenon is perceived as difficulty in being able to hear elements of the target speech signal. The other one is high-level 'informational' masking that occurs when both target and interference share *similar* sounds, which do not overlap energetically (and can be separately heard) but are nevertheless difficult to 'disentangle'. In this case, a listener is able to hear the components of each source but finds it difficult to select out the components belonging to the attended source. In everyday situations, both these types of masking will be present. In overlapping speech conditions, informational masking can become particularly problematic. Although ASR systems may work quite differently from the human speech perception system, they are still faced with both these types of masking effect. Designing solutions that work well when both are present (e.g. in overlapping speech) is a challenge.

To improve the robustness of an ASR system to overlapping speech, speech separation techniques can be employed as a distinct pre-processing stage [Yoshioka et al., 2018a]. A common pipeline is that a speech separation system processes a mixture signal to segregate individual speech sources, from which target speech is selected and decoded by an ASR system. Two facts make the speech separation by machine possible. One is that a person's voice contains unique spectral characteristics, i.e. the distribution of frequencies of which the sound is compressed. And the other is that source spatial locations differ. However, exploiting these 'cues' in real situations is not straightforward.

Speech separation has been an active research topic in signal processing for decades. Even when good spectral or spatial cues exist, separation is challenging and the speech output of separation systems will contain residual noise from the interferer and possibly further distortions and artifacts produced by the separation processing itself. Deep-learning based models, namely deep neural networks, provide a powerful modelling capacity and have revolutionised speech separation systems to achieve unprecedented separation accuracy [Hershey et al., 2016; Kolbæk et al., 2017]. Most speech separation methods operate on time-frequency domain signals, which are obtained by applying a short time Fourier transform (STFT) to the time-domain signals. Recently, end-to-end separation approaches that directly process time-domain signals have made large progress and achieved a new state-of-the-art performance for single-channel separation [Luo and Mesgarani, 2019]. Specifically, neural networks have been demonstrated to be able to replace the conventional STFT operation. However, the single-channel separation system is not robust to noise and reverberation, which commonly exist in audio signals recorded by distant microphones [Heitkaemper et al., 2020].

Another approach to speech separation is to exploit 'spatial cues'. Nowadays, it is common to have a set of microphones arranged in some fixed geometry which work together, which is called a microphone array. Multi-channel recordings from a microphone array provide spatial information related to source location, which has been shown to increase separation system robustness to reverberation and noise [Wang et al., 2018]. However, most multi-channel separation approaches focus on time-frequency domain processing, and multi-channel recordings have not been exploited for the recent successful end-to-end signal processing framework. The question of how best to develop end-to-end time domain approaches for multi-channel recordings remains an unsolved problem.

1.2 Aim and Objectives

The aim of this thesis is to re-address the problem of distant speech recognition in noisy multi-talker scenarios by exploiting multi-channel recordings and recent advances in deep-learning techniques. The thesis will start by making an analysis of real conversational speech recordings to better understand the specific challenges. It will then pursue three specific questions that are introduced below, which will be expanded in later chapters. These concern how best to exploit multiple channels in the design of deep-learning based separation systems; how to target source separation systems towards specific speakers; and finally, how to train separation systems in situations where little or no unmixed data is available.

Analysis of real conversational speech recordings

Early studies of distant microphone speech recognition have usually overlooked the overlapping speech problem when researchers designed distant speech recognition systems for natural conversations. This is partially due to the fact that most corpora are recorded in meeting scenarios where the proportion of overlapping speech is small, i.e. around 10% [Cetin and Shriberg, 2006]. However, it was found that overlapped segments in meetings are hard to recognise [Renals and Swietojanski, 2017]. The current state-of-the-art ASR achieves a performance around 30% in terms of word error rate (WER), when recognising conversations recorded by using distant microphones in meetings [Yoshioka et al., 2018b]. Lately, a study in Barker et al. [2018] showed that conventional ASR systems that are developed for single-talker speech failed to recognise conversations happening in multi-talker dinner party scenarios. The achieved WER using distant recordings is 81%, compared to 48% achieved using close-talk microphones, showing that there is a large performance gap. The overlapping speech has been recognised as a main cause of the large gap. However, there is limited understanding of the extent of overlapping speech in casual conversations and how overlapping speech impacts on ASR. This knowledge is necessary to guide researchers to develop techniques such as speech separation to improve distant speech recognition in multi-talker scenarios.

Exploiting multi-channel recordings

Conventional speech separation methods operate on time-frequency domain signals. Time-frequency representations are obtained by applying a STFT to the time-domain signals with a window length around 30-100 ms. The time-frequency representation is constructed from complex numbers containing a real part and an imaginary part, which can be expressed in terms of magnitude and phase in the polar coordinate system. The majority of traditional separation methods only modify the magnitude and discard the phase information. The inattention to phase information in the early days was mainly due to it being believed that using noisy phase information does not cause much perceptual loss, as long as magnitude information can be reliably recovered [Wang and Lim, 1982]. Later, however, it was shown that speech perceptual quality can be improved by only enhancing phase without changing magnitude information [Paliwal et al., 2011]. Although there are recent studies enhancing phase information to achieve better quality of reconstructed signals [Williamson et al., 2016], it was argued that separate real and imaginary components in complex time-frequency representations may not be optimal for the speech separation task [Luo and Mesgarani, 2019].

A natural way to take into account both the magnitude and phase information together during separation is directly modelling time-domain signals [Luo and Mesgarani, 2018b]. The time-domain approach can jointly optimise magnitude and phase components without splitting them into different streams. Instead of transforming time-domain signals to the time-frequency domain, a recently proposed end-to-end time-domain approach used neural networks to learn representations from time-domain signals [Luo and Mesgarani, 2019]. Then, the separation process operated on the learned representations that equalled to jointly optimising the magnitude and phase information. The end-to-end method achieved better separation performance compared to time-frequency approaches with optimal magnitude information.

With the development of commercially available devices mounted with a microphone array, multi-channel recordings are commonly used for speech separation [Wang et al., 2018; Yoshioka et al., 2018a]. Spatial features that relate to source locations can be extracted from multi-channel recordings and used as additional input to a separation network to provide spatial information. Frequency-domain spatial features have been demonstrated to be beneficial for multi-channel separation on the time-frequency domain, especially when sources arrive from different angles [Wang et al., 2018; Wang and Wang, 2018]. However, time-frequency spatial features cannot be easily incorporated in the end-to-end time-domain separation framework. There are two reasons for this. The first is that the frequency-domain spatial features cannot be perfectly aligned with the spectral features learned from time-domain signals using trainable layers, since the window lengths required to extract the two types of features are different. It has been demonstrated that simply concatenating asynchronous features from different streams suffers in performance [Wölfel and Woelfel, 2009]. Second, there will be a mismatch between the spatial and spectral representations when they are extracted from signals on different domains. However, using neural networks to extract spatial features from time-domain signals seems to be a potential solution, since the spectral features can be learned with trainable layers in the end-to-end single-channel separation system [Luo and Mesgarani, 2019]. Therefore, this thesis will investigate how to use neural networks to extract spatial features from time-domain multi-channel signals for speech separation.

Modelling raw audio is difficult because signal samples in the time-domain change quickly and speech patterns exist within both short and long time scales [van den Oord et al., 2016]. Neural networks can capture long temporal dependencies and have been successfully applied to monaural speech separation [Luo and Mesgarani, 2018b, 2019]. The design of separation network architectures has a significant impact on the separation performance. Since the speech signal is a temporal sequence, a model should be able to learn both long-term trends while preserve short-term details. Many studies explore strategies to increase the temporal modelling capacity of the network to improve the single-channel separation performance, including using long short-term memory (LSTM) networks [Luo and Mesgarani, 2018b], temporal convolutional network (TCN) [Luo and Mesgarani, 2019], and splitting a long sequence into short chunks to model a long sequence signal more efficiently [Luo et al., 2020]. But it is unclear if the achievement in the single-channel separation in a theoretically anechoic condition could benefit the multi-channel separation in a noisy environment.

Exploiting speaker identity information for speech separation

A speech separation system separates as many sources in a mixture as possible. However, in practice, we are only interested in recognising speech from specific speakers instead of all. Although a speaker recognition system can be used to select speech signals belonging to a target speaker from all separated signals, it is not optimal because the selection process will introduce extra errors [Delcroix et al., 2020]. In addition, the design of a separation system requires the prior knowledge of number of speakers in a mixture. Once a separation system is trained with a fixed number of speakers, it is inconvenient to apply this separation system to scenarios in which the number of active speakers could vary in different time periods.

Speaker identity information has been exploited in a separation system design to focus its attention on specific speakers, which is known as 'informed extraction' [Wang et al., 2019b]. The speaker identity information that can be represented as voice characteristics inform a speaker extraction system to extract a target speaker owning the same voice [Wang et al., 2019b;

Žmolíková et al., 2019]. An extraction system has three main advantages over a separation system. First, it does not need to know the total number of speakers in a mixture and can be applied to mixtures containing various number of sources. Second, the speaker identity of output speech from an extraction system can be controlled by the input speaker identity information. Last, since the speaker identity information provides additional prior knowledge of target speakers, the knowledge has been demonstrated to improve single- and multi-channel extraction systems [Delcroix et al., 2018; Žmolíková et al., 2017].

Therefore, this thesis will investigate how to exploit speaker identity information in an end-to-end time-domain multi-channel extraction system. Recently, single-channel speech extraction systems have employed the end-to-end time-domain framework and achieved better performance compared to time-frequency domain systems [Delcroix et al., 2020; Xu et al., 2020a]. The time-domain extraction system can be further improved when multi-channel recordings are available. A common approach is to use frequency-domain spatial features as additional input to the extraction system [Delcroix et al., 2020]. However, as already mentioned, frequency-domain features may not be optimal for the end-to-end framework. This raises a question for the extraction system design concerning how to input and fuse features from different modalities within the extraction network.

Unsupervised and semi-supervised training for separation networks

The common framework used for training a separation network is supervised training, which requires large amounts of paired mixture and reference data. A well known problem of supervised learning is its poor generalisation ability to unseen scenes. Currently, most deep learning based speech separation systems are developed under supervised training with simulated data, in which a mixture signal is artificially generated from single-speaker speech signals processed by existing simulation tools. A separation system trained on simulated data usually performs poorly with real data, due to the mismatched conditions between the real and simulated environments. With the existing simulation tools, it is difficult to fully capture the acoustic characteristics of a real environment and generate data that matches real recordings. This is because, in the real environment, the distribution of sound types and acoustic conditions such as reverberation may be unknown and hard to estimate.

Intuitively, using data collected in real environments to train a separation model can potentially solve the mismatch problem. However, in real environments, it is hard or even impossible to conduct supervised training for a separation system because paired mixture and clean sources cannot be easily obtained, and only large amounts of unlabelled mixtures are available. Unsupervised and semi-supervised learning approaches explore strategies to use unlabelled data and provide potential solutions to establish a deep learning based separation system in real environments. A recent unsupervised separation approach generates artificial mixtures of unlabelled mixtures, which are used as input to a separation system during training. The separation model is trained to estimate and assign sources back to original unlabelled mixtures [Wisdom et al., 2020]. However, this unsupervised approach still suffers a mismatch problem between training and testing since the artificially created mixtures contain more sources than real mixtures [Tzinis et al., 2021]. This mismatch causes a large performance gap between unsupervised and supervised approaches.

In summary, this thesis aims to address the following research questions listed as below.

- What is the extent of overlapping speech in casual conversations and what impacts can overlapping speech have on current state-of-the-art ASR systems?
- How can multi-channel information be effectively exploited in an end-to-end time-domain speech separation system?
- How can speaker identity information be exploited in an end-to-end time-domain multi-channel extraction system?
- How to build a speech separation network in situations where the amount of supervised training data is limited?

1.3 Contributions

The contributions from the work are listed as following:

1. The first contribution is a thorough analysis about the extent of overlapping speech in a natural multi-talker party scenario. The analysis shows that the speaker overlaps occur

in up to 30% of whole conversations in a multi-talker dinner party scenario and provides evidence that the overlapping speech has a significant negative impact on the performance of distant speech recognition. As an initial attempt, an existing speech separation technique is applied as a front-end to address the problem caused by overlapping speech. The speech separation method provides a substantial improvement compared with the unprocessed signal, which indicates that the separation technique is the key to distant speech recognition with overlapping speech. This piece of work has formed part of work published at the ISCA CHiME 2018 workshop [Xiong et al., 2018].

2. An end-to-end time-domain multi-channel separation system has been designed for addressing overlapping speech recognition in noisy and reverberant environments. The system employs a trainable 2-D convolutional layer to extract spatial features from pairs of microphone signals. The learned features better suit the time-domain multi-channel separation system than the hand-crafted time-frequency features and improve both speech separation and recognition performance. The design of separation model architectures has been explored for better use of the multi-channel information. Increasing the temporal modelling capacity of a multi-channel separation network improves the separation performance. In addition, the influence of reverberation on the separation task has been investigated, showing that the end-to-end approach is sensitive to reverberation. However, applying a dereverberation method as a pre-processing stage can reduce the effect of reverberation and further improve the speech recognition performance. Last, a multi-stage speech separation framework is introduced to enhance the signal quality of separated speech. This work has led to a conference publication at ICASSP 2020 [Zhang et al., 2020a].
3. The next piece of work studies how to incorporate knowledge of speaker identities into a time-domain multi-channel extraction system. The developed end-to-end multi-channel separation system is extended to an extraction system. To efficiently combine spectral features, spatial features, and speaker identity features, a novel feature fusion strategy is proposed. Informed by identity information of target speakers, the system is able to simultaneously identify and output corresponding sources from a noisy and reverberant mixture. This piece of work has formed the basis of a paper that has been accepted for publication at ICASSP 2021 [Zhang et al., 2021b].
4. The final contribution studies how to apply an end-to-end separation system to situations

where the amount of supervised training data is limited. This work first shows that existing unsupervised approaches suffer a mismatch problem between training and testing data. To address the mismatch problem, a novel unsupervised approach is proposed, which combines a teacher-student learning approach and a remixing strategy. In addition, to improve the unsupervised trained model in situations where a small amount of ground-truth data is available, a semi-supervised approach is developed, which exploits fine-tuning and knowledge distillation strategies. This piece of work has formed the basis of a paper that has been accepted for publication at INTERSPEECH 2021 [Zhang et al., 2021a].

1.4 Thesis Outline

The remainder of the thesis is structured as follows:

- Chapter 2 provides a review of progress that has been made in recent years on the speech recognition in multi-talker environments. It also presents a critical review of deep learning based speech separation approaches in the literature.

- Chapter 3 presents analysis of overlapping speech in natural conversations in an informal social scenario. It shows what impacts the overlapping speech can have on the performance of an ASR system. An existing speech separation method is applied as an initial attempt to solving the overlapping speech problem.

- Chapter 4 investigates the design for neural network based time-domain multi-channel speech separation.

- Chapter 5 studies how to exploit speaker identity information in a time-domain multi-channel extraction system to extract target speakers from a mixture.

- Chapter 6 studies unsupervised and semi-supervised learning approaches for end-to-end multi-channel speech separation in noisy environments.

- Chapter 7 summarises the main findings of the dissertation and presents potential directions for future research.

Chapter 2

Multi-talker speech recognition and separation

Distant speech recognition in everyday environments is a challenging task for machines. This is because many factors in daily environments can degrade the quality of a speech signal, including ambient background noise, reverberation, and interfering speech from competing speakers. In order to facilitate the study of distant speech recognition, research data in various environments have been collected and released [Barker et al., 2015, 2018; Kinoshita et al., 2013]. Many methods have been developed to improve the robustness of distant speech recognition systems and evaluated with the publicly available data [Barker et al., 2017; Vincent et al., 2017]. These studies have shown that the effects of noise and reverberation on an ASR system can be significantly reduced by using powerful denoising and dereverberation methods. However, recognising multi-talker speech remains an unsolved problem. This can be observed when attempting to recognise speech recorded in multi-party meeting scenarios, where the current state-of-the-art word error rates (WERs) are still around 30% [Yoshioka et al., 2018b]. In real homes, ambient noise, reverberation, and overlapping speech usually exist at the same time, causing many difficulties to current speech recognition systems [Barker et al., 2018].

To overcome the problem caused by overlapping speech, one of the popular approaches is to apply a speech separation technique as a pre-processing step for ASR. Many efforts have been dedicated to developing suitable speech separation techniques and the field of multi-talker speech separation has changed dramatically during the last couple of decades, progressing from signal processing methods to deep learning supported approaches. The powerful modelling capacity of neural networks has boosted separation performance by a large margin compared to the traditional signal processing approach. Depending on the number of microphone channels used in a separation system, algorithms can be classified into two groups, namely single-channel separation and multi-channel separation. In both cases, deep learning approaches are now heavily employed [Wang and Chen, 2018]. In the single-channel case, methods take advantages of the difference of the spectral characteristics between different sources for separation. While for the case of multi-channel recordings, the focus is on how to exploit the spatial information for the separation task.

Speech separation techniques can be further categorised into two main classes, speaker independent separation and speaker dependent separation. Speaker independent separation (or blind speech separation) separates all speech sources from a mixture signal. There is no need of prior knowledge of the speaker identity for such a system but prior knowledge of the number of sources in a mixture is usually needed for the system design. In contrast, speaker dependent separation or speaker extraction systems separate specific speakers from a multi-talker mixture [Žmolíková et al., 2019]. This selective attention can be obtained by forcing the system to learn the characteristic of a specific speaker or it can be informed by external bias that relates to the speaker identity. One advantage of the speaker extraction system over speaker-independent separation system is that speaker extraction systems requires no prior knowledge of the number of speakers in a mixture.

The other main categorisation is whether systems are trained in a supervised or an unsupervised manner. Most recent deep learning based speech separation systems operate under the supervised training framework. The supervised learning framework requires pairings of noisy mixtures and the associated clean sources contained in the mixture to train a system. However, a widely known problem of supervised training is its poor generalisation to conditions that are mismatched to the training data. This means that supervised systems only work in a limited way when applied to real environments. To address the problem caused by mismatched conditions, unsupervised and semi-supervised learning approaches have been proposed to use noisy signal from the target domain to adapt a separation model [Lam et al., 2020; Wisdom et al., 2020].

There have been many previous attempts to address the problems described above. This chapter reviews this literature focusing on works that are related to or employed in the later chapters of this thesis. Section 2.1 reviews the challenges that have been faced during tackling the distant speech recognition and the corresponding progress. The review is conducted from the perspective of speech corpora that have been designed to target this problem. Section 2.2 describes the general framework of the speech separation task for the case of a single microphone. It covers the basic principles and detailed description of supervised speech separation approaches based on deep learning. Section 2.3 extends the previous section to cover the development of speech separation techniques when multiple microphones are available. Section 2.4 focuses on the review of speaker extraction systems that exploits additional speaker information for separation. Section 2.5 discusses unsupervised learning techniques to train speech separation in situations where clean sources are not accessible.

2.1 Progress and challenges of multi-talker speech recognition

To facilitate the development of distant speech recognition systems, many corpora have been developed for various application scenarios. Widely used standard databases include AMI [Renals et al., 2008], CHiME series [Barker et al., 2015, 2013, 2018], and REVERB [Kinoshita et al., 2013].

Automatic speech recognition has achieved good performance in read speech corpora such as Wall Street Journal (WSJ) [Paul and Baker, 1992] and Librispeech [Panayotov et al., 2015]. The current state-of-the-art ASR system can achieve recognition performance of less than 5% in terms of WER for these read speech corpora [Gulati et al., 2020]. In daily life, conversational speech occurs more commonly compared to read speech or prepared speech [Tucker and Ernestus, 2016]. However, recognising conversational speech in everyday environments is still challenging and the performance falls far behind that of prepared speech recorded in studios. The extra difficulty comes from the casual speech style, non-stationary ambient noise, highly reverberant environments, and competing speech sources.

In recent years, speech corpora have been increasingly focused on more challenging tasks that better approximate scenarios in the real world. The recognition of conversational speech starts from scenarios of meetings and lectures. For instance, ICSI [Janin et al., 2003], CHIL [Mostefa et al., 2007], and AMI [Renals et al., 2008] are meeting corpora recorded from academic laboratories using expensive recording devices that are not consumer grade. An analysis using the AMI corpus shows that there is a big performance gap between speech signals coming from close-talk microphones and distant microphones [Renals and Swietojanski, 2017]. Even if using multi-channel speech enhancement techniques such as beamforming and a strong CNN model can improve the performance in the distant case, the achieved WER (46.8%) is much worse than the WER observed in the close talking case (25.6%). Renals and Swietojanski [2017] suggested that overlapped speech is a main factor that cause this difference.

For conversations recorded in meetings, the ratio of speech overlapped by one or more talkers is around 12% [Cetin and Shriberg, 2006]. Cetin and Shriberg [2006] also demonstrated that the overlapped speech affects an automatic speech recognition system not only by acoustic cross-talk, but also by the differences in speech style and content near overlapped regions compared to speech elsewhere. However, researchers did not pay much attention to the issue of multi-talker speech recognition at that time because of the relatively small ratio of overlapped speech in the meeting corpora being studied.

Overlapping speech recognition in daily environments has drawn a lot of attention in recent years. The PASCAL speech separation challenge aims to recognise speech in artificial speech-in-speech mixtures [Cooke et al., 2010]. The task is designed to recognise keywords from a utterance spoken by a target speaker when artificially mixed by a similar sentence from another competing speaker. Due to the simplified speech style, the submitted systems achieved high keyword recognition accuracy and some of the systems even outperformed human performance [Hershey et al., 2010]. Building on this earlier PASCAL speech separation challenge, the PASCAL CHiME challenge was the first attempt to link the speech separation techniques to the application of speech recognition for overlapped speech in real everyday living environments [Barker et al., 2013]. The speech was artificially mixed with backgrounds recorded by a two-channel microphone in a domestic environment. However, these corpora only consider the case of small scale vocabulary size. Therefore, the following CHiME-2 [Vincent et al., 2013] increased the vocabulary size to a medium size by using the Wall Street Journal corpus that is constructed from a 5k vocabulary read speech.

Considering that CHiME-1 and CHiME-2 simulated mixtures using either a fixed or slowly varying impulse responses, they were not considered sufficiently realistic for drawing conclusions about real scenarios. To directly compare simulated mixtures with mixtures recorded directly in real noisy environments, the CHiME-3 recorded read speech using a 6-channel tablet based microphone array in noisy environments [Barker et al., 2015]. The speaker held the device and spoke prepared sentences from the WSJ0 corpus in four varied environments: cafe, street junction, public transport and pedestrian area. The corpora provided a simulation tool to approximate the characteristics of real conditions. REVERB [Kinoshita et al., 2013] was designed to investigate how reverberation affects quality of speech signals and the ASR performance. The REVERB corpus provided simulated reverberant audios and real recordings in office rooms for comparison. These challenges study a mismatch case that various techniques are trained on simulated data while evaluated on both simulated and real test data.

Although the CHiME-3 dataset is recorded in real-world noisy environments, the interfering sources consist of only background noise and there is rarely competing speakers. In addition, the speech is from prepared sentences in a reading corpus instead of spontaneous speech. There are a few real large-scale recordings for the task of spontaneous speech recognition: Santa Barbara Corpus of Spoken American English [John W. Du Bois and Thompson, 2000], Sheffield Wargame Corpus [Fox et al., 2013] and CHiME-5 corpus [Barker et al., 2018]. Using single-talker speech recognition systems achieved poor performance in these datasets. One of the difficulties is the large proportion of overlapped speech that occurs in less formal multi-talker scenarios.

The CHiME-5 [Barker et al., 2018] dataset, designed for robust ASR evaluation, recorded speech of natural conversations between four talkers who are friends to each other in home environments. The natural conversations between participants caused an utterance to be partially or fully overlapped by other competing utterances. The audio is recorded using both near-field microphones worn by each participant, and commercially-available distant microphone arrays positioned near a wall or at a corner of a room to mimic real home settings. It was observed that the SNR of the target speech to the interfering speech in the distant recordings is much lower than the speech recorded using near-field microphones. The baseline system provided in Barker et al. [2018] used the state-of-the art acoustic model at that time and achieved 81.3% WER for distant microphone recordings in the dev set. The performance of distant recordings is much worse than that achieved by near-field microphones which is around 40%. Although the distant speech recognition performance has been improved to 36.9% of WER by current state-of-the-art system [Medennikov et al., 2020], it is still a lot higher than the best performances for Librispeech and other read speech corpora. The poor performance highlights that recognising overlapping speech recorded by distant microphones in noisy environments remains unsolved.

2.2 Single-channel speech separation

Public speech corpora have facilitated the development of distant speech recognition systems in various environmental conditions. Among all successful methods, front-end processing is usually shown beneficial to improve distant speech recognition in noise [Heymann et al., 2016] and in reverberation [Kinoshita et al., 2016]. Speech separation techniques have been applied as a front-end processing stage to speech recognition in multi-talker scenarios, attempting to address problems caused by overlapping speech [Boeddecker et al., 2018].

Speech separation techniques have been widely studied for decades, ranging from signal processing based methods to recent deep learning based approaches. This section aims to review the most relevant recent developments, but for a comprehensive account the reader is directed to Wang and Chen [2018]. This section first introduces widely used corpora for evaluating speech separation techniques. The second part describes the physical formulation of the speech separation problem and reviews deep learning based single-channel speech separation techniques.

2.2.1 Corpora for speech separation

To facilitate the development of the speech separation technique, several corpora based on simulation have been released. Simulated data allows researchers to access oracle clean signals to build supervised speech separation systems and allows separation systems to be measured by signal quality based measurement such as signal to distortion ratio (SDR) [Vincent et al., 2006]. A summary of commonly used simulated data corpora is shown in Table 2.1.

One of the currently most widely used speech separation datasets is WSJ0-2mix [Hershey et al., 2016]. WSJ0-2mix uses clean read speech from the earlier published Wall Street Journal (WSJ0) dataset [Paul and Baker, 1992] to simulate 2-speaker mixtures by randomly selecting two utterances and summing them with pre-defined scale factors [Hershey et al., 2016]. This simulation results in a signal-to-noise ratio (SNR) uniformly distributed between -5 and 5 dB. WSJ0-2mix provides training, validation, and test data, each containing 20,000, 5,000, and 3,000 utterances, respectively. The training set and validation set in WSJ0-2mix are generated from the WSJ0 si_tr_s set and the test set is generated from the WSJ0 si_dt_05 and si_et_05 sets. All the mixtures contain a 100% overlap ratio between two speakers and are assumed to be in a clean condition without noise and reverberation.

To develop speech separation techniques in realistic conditions including noise and reverberation components, spatialised WSJ0-2mix [Wang et al., 2018], WHAM! [Wichern et al., 2019], and WHAMR! [Maciejewski et al., 2020] have been proposed. The spatialised WSJ0-2mix extends the WSJ0-2mix to a reverberant version by convolving the clean speech signals with artificial room impulse responses (RIRs) generated by using simulation tools [Allen and Berkley, 1979]. The WSJ0 Hipster Ambient Mixtures (WHAM!) dataset [Wichern et al., 2019] extends the WSJ0-2mix to a noisy version by mixing speech mixtures with pre-recorded real world noise. To better approximate realistic conditions, WHAMR! dataset [Maciejewski et al., 2020] extends the WHAM! dataset with artificial reverberations.

The above mentioned datasets are created with a 100% overlap ratio, which is severely mismatched to the overlap ratio in a natural conversation. To mimic realistic scenarios, LibriMix [Cosentino et al., 2020] and SMS-WSJ [Drude et al., 2019b] have simulated partially overlapped speech. LibriMix [Cosentino et al., 2020] uses speech utterances in the Librispeech dataset to simulate mixture data and contains a partially-overlapped test set in which the overlap ratio between speakers are uniformly distributed between 0% and 100%. SMS-WSJ [Drude et al., 2019b] simulated a reverberant data with overlap ratio ranging from 0% to 100%, following a normal distribution.

The LibriCSS dataset [Chen et al., 2020b] is designed for the continuous speech separation and recognition task. This corpus contains 10 hours of distant speech recordings designed for evaluating speech separation algorithms. The audio is recorded by using loudspeakers to play back concatenated LibriSpeech utterances in a regular meeting room. The loudspeakers are randomly positioned in the room and a seven-channel circular microphone array positioned in the centre area of the room is used to capture the signals. The overlap ratios of mixture signals are controlled to fall in the range between 0% and 40%.

Table 2.1: The most commonly used corpora for recent speech source separation research

Dataset	Origin data	Reverberation	Noise	# microphones	hours (training)	# speakers (training)	Overlap style
WSJ0-2mix [Hershey et al., 2016]	WSJ0	None	None	1	30	101	Full
WHAM! [Wichern et al., 2019]	WSJ0	None	Recorded	1	58	101	Full
WHAMR! [Maciejewski et al., 2020]	WSJ0	Image method	Recorded	2	58	101	Full
SMS-WSJ [Drude et al., 2019b]	WSJ0&1	Image method	White Noise	6	-	101	Full/Partial
LibriMix [Cosentino et al., 2020]	LibriSpeech	None	Recorded	1	270	1172	Full / Partial
LibriCSS [Chen et al., 2020b]	LibriSpeech	Recorded	Recorded	7	10	40	Partial

2.2.2 Time-frequency domain speech separation

In the previous section, common datasets for developing and evaluating speech separation techniques have been described. This section will introduce the problem formulation of speech separation and then review existing deep-learning based speech separation techniques. Assume that a microphone observes an audio signal $y(t)$ generated by source speech signals $\{s_c\}_{c=1}^{C}$ from C speakers and independent background noise n. It can be formulated as follows.

$$y(t) = \sum_{c=1}^{C} h_c(t) * s_c(t) + n(t) \tag{2.1}$$

where $h_c(t)$ denotes the room impulse response between the source s_c and the receiver and $*$ is a convolution operation. Due to the limited speech information contained in one time sample in a time-domain signal, representations extracted from multiple samples are commonly used in speech processing. One popular representation is the time-frequency (T-F) representation which is calculated by applying short-time Fourier transform (STFT) to the time-domain signal. Each frame has a duration of 20 ms. On the frequency domain, Equation 2.1 turns into follows.

$$Y(k, f) = \sum_{c=1}^{C} H_c(f)S_c(k, f) + N(k, f) \tag{2.2}$$

where $Y(k, f)$, $S_c(k, f)$, and $N(k, f)$ denote the T-F bin of the observed signal, the c-th source signal, and the noise received at frame k and frequency f. And $H_c(f)$ denotes the Fourier transform of $h_c(t)$. The aim of time-frequency domain speech separation is to recover the individual sources $\hat{S}_c(k, f)$ given a mixture signal. Single-channel speech separation techniques use the mixture signal recorded by a single microphone to estimate the sources.

Deep learning based methods exploit the powerful modelling capacity of neural networks and directly learn the mapping from a noisy input to source signals. The input to a network can use the time-frequency representations or other features extracted from the T-F representations such as square magnitude $|Y(k, f)|$, Mel-frequency filterbank [Han et al., 2015], and log Mel-frequency filterbank [Xu et al., 2015].

A speaker separation model can be categorised as either a speaker-dependent model or a speaker-independent model, depending on whether speakers are required to be the same or not between training and testing. There are two disadvantages in a speaker-dependent model. One is that the speaker-dependent model is not flexible to deal with new speakers. Every time a new target speaker is to be separated, a new speaker-dependent model should be trained on this specific speaker. The other is that training a speaker-dependent model requires large amounts of audio data from each target speaker. Therefore, it is desired to have a speaker-independent separation system that can generalise well to unseen speakers during training. However, when building a deep-learning based speaker-independent separation model, it is difficult to assign the model output to its corresponding speaker since the order of model outputs is arbitrary. It will be hard for a neural network to learn a mapping from a mixture to individual sources if the assignment is wrong. This is known as the arbitrary permutation problem [Hershey et al., 2016]. Clustering based methods [Hershey et al., 2016] and the Permutation Invariant Training (PIT) technique [Yu et al., 2017] have been proposed to alleviate this issue.

Deep clustering (DC) [Hershey et al., 2016] learns an embedding vector representing each T-F bin by using a neural network. The neural network is trained in a way such that the representations belonging to the same source are drawn closer together and the representations generated from different sources are pushed further apart. Then, a K-means clustering algorithm groups together the representations belonging to the same source and assigns the segmentation label to each T-F unit. The deep clustering loss is formulated as follows.

$$\mathcal{L}_{\mathrm{DC}}(V, U) = \|VV^T - UU^T\|_F^2 \tag{2.3}$$

where V and U indicate the learned embedding vector and the true cluster labels, respectively. The deep clustering model has been shown to effectively serve as a front-end processing for the speech recognition task in multi-talker scenarios [Isik et al., 2016; Menne et al., 2019]. Using a GMM-based acoustic model, the deep clustering source separation front-end led to a WER of 30.8% on WSJ0-2mix, compared to 89.1% for the unprocessed mixtures [Isik et al., 2016].

Similar to deep clustering, a deep attractor network [Chen et al., 2017] learned high dimensional embedding representations for time-frequency units of acoustic signals. Instead of using the K-means algorithm, the deep attractor network created a reference point for each source, which determined the similarity of each T-F bin in the mixture to each source. The T-F bins dominated by a specific source will be pulled together to the corresponding reference point.

However, these clustering approaches cannot be trained to directly map a noisy signal to individual source signals. In order to train a separation network to directly learn the mapping between the mixture and the sources, permutation invariant training PIT [Yu et al., 2017] addresses the permutation problem by dynamically calculating the training loss function. It firstly calculates loss between reference features $\mathbf{S} = \{S_c\}_{c=1}^{C}$ and estimated features $\hat{\mathbf{S}}$ for all possible assignments P. Then, the minimum loss is selected and used for back-propagation to update the parameters of the separation network. The PIT loss $\mathcal{L}_{\text{PIT}}$ can be formulated as below.

$$\mathcal{L}_{\text{PIT}}(\mathbf{S}, \hat{\mathbf{S}}) = \min_{\mathbf{P}} \sum_{c=1}^{C} \mathcal{L}(S_c, [\mathbf{P}\hat{\mathbf{S}}]_c) \tag{2.4}$$

where $\mathbf{P}$ is a permutation matrix and $\mathcal{L}$ is a signal-level loss function to be minimised. To benefit from both the embedding generation task and the signal mapping task, a multi-task learning framework is proposed, namely ChimeraNet, which combines both deep clustering loss and PIT loss [Luo et al., 2017]. This system has two output heads, one for estimating embeddings for clustering, the other one for the PIT training.

Depending on the form of training targets, deep learning based methods can be categorised into masking based methods or mapping based methods. For the masking based method, widely used training targets include ideal binary mask (IBM) [Wang, 2005], ideal ratio mask (IRM) [Srinivasan et al., 2006], phase sensitive mask (PSM) [Erdogan et al., 2015] and complex ideal ratio mask (cIRM) [Williamson et al., 2016]. The IBM is defined on the time-frequency representation of a noisy signal based on the signal-to-noise ratio (SNR), such as

$$IBM(k,f) = \begin{cases} 1, & if\, SNR(k,f) > threshold \\ 0, & otherwise \end{cases} \tag{2.5}$$

The equation 2.5 shows that IBM will be set as 1 when the time-frequency bin belongs to the related source and also maintains a good signal quality. The IRM is a soft version of IBM. It is ranged from 0 to 1, representing the probability that the time-frequency unit belongs to each source. The definition of IRM is given as:

$$IRM(k,f) = \frac{|S(k,f)|^{\alpha}}{|S(k,f)|^{\alpha} + |N(k,f)|^{\alpha}} \tag{2.6}$$

where α is a warping factor of the magnitudes to control the sharpness of the mask, which is commonly chosen as 0.5.

Most of methods operating on the time-frequency domain only use magnitude information of the T-F bins and discard the phase information due to the belief that noisy phase information does not cause much perceptual loss, as long as magnitude information can be reliable recovered [Wang and Lim, 1982]. However, in environments with strong noise and room reverberation, directly using the phase from the observed signal to reconstruct estimated sources will cause phase inconsistency issues [Williamson et al., 2016]. To involve the phase information in the mask, PSM [Erdogan et al., 2015] has been proposed, which aims to maximise the signal-to-noise ratio while restricting the mask to a real value:

$$PSM(k,f) = \frac{|S(k,f)|}{|Y(k,f)|} cos(\angle S(k,f) - \angle Y(k,f)) \tag{2.7}$$

where $(\angle S(k,f) - \angle Y(k,f))$ denotes the phase difference between the source and the observed signal. An alternative approach to consider phase information is to directly operate in the complex time-frequency domain. Motivated by this, a complex ideal ratio mask (cIRM) has been proposed, which contains both real and imaginary components [Williamson et al., 2016].

The mapping based method aims to directly estimate source signals from a noisy observation. It has been shown to avoid the corruption caused by reconstruction process in the mask based method [Huang et al., 2014]. Mean Squared Error (MSE) is usually used as loss function that can be expressed as:

$$MSE(k,f) = (\hat{M}(k,f)|Y(k,f)| - |S(k,f)|)^2 \tag{2.8}$$

where $\hat{M}(t,f)$ denotes estimated source masks. The mapping based system can be trained in a progressive way [Weninger et al., 2014]. In the first stage, the system is trained to estimate the mask from a noisy signal. Then, additional layers could be added to the trained structure in the first stage and further optimised with the MSE loss function. More recently, to improve the enhancement performance, researchers have considered involving phase information in the mapping based method. Wang et al. [2020] has proposed a complex spectral mapping loss function reconstruct both the real and imaginary components.

2.2.3 Time domain speech separation

It was argued that separate real and imaginary components in complex time-frequency representations may not be optimal for the speech separation task and most time-frequency domain systems have not considered recovering phases for individual sources [Luo and Mesgarani, 2019]. An alternative approach to jointly using magnitude and phase information is directly modelling the time-domain signal. Several time-domain approaches have been proposed for the single-channel speech separation task, one of which is Conv-TasNet [Luo and Mesgarani, 2018b, 2019]. The Conv-TasNet takes the raw waveform mixture as input and outputs waveform signals for each source. This system replaces the fixed STFT and its inverse with trainable layers, i.e., 1D convolutional layers, which are trained directly to process the time-domain signals. The 1D convolutional kernel moves along one dimension to calculate the output, which is the time-axis in a speech signal. The time-domain approach directly optimises a signal reconstruction objective function on the time-domain, such as scale-invariant signal-to-noise ratio (SI-SNR), which has been shown to achieve good general performance across a range of speech enhancement evaluation metrics [Kolbæk et al., 2020]. The SI-SNR is expressed as:

$$\begin{aligned}\mathcal{L}^{\text{SI-SNR}} &= 10\text{log}_{10}\frac{||\text{s}_{target}||^2}{||\text{e}_{noise}||^2}\\ \text{s}_{target} &= \frac{\langle\hat{s}, s\rangle s}{||s||^2}\\ \text{e}_{noise} &= \hat{s} - s_{target}\end{aligned} \tag{2.9}$$

where $\hat{s}$ and s denote an estimated source and a clean source, respectively, $\langle\hat{s}, s\rangle$ denotes the inner product between $\hat{s}$ and s, and $||s||^2$ denotes the signal power.

As illustrated in Figure 2.1, the overall structure uses an encoder-decoder framework. A separation block is built to estimate masks for each source, which are multiplied on the encoded representation to generate separate representations for each source. The separation block can be constructed by recurrent neural networks, i.e. LSTM [Luo and Mesgarani, 2018b] or temporal convolutional networkss (TCNs) [Lea et al., 2016].

A TCN stacks multiple convolutional layers into a hierarchy to model temporal signals with long sequence lengths. The TCN has been shown to be capable of capturing long-range dependencies while significantly reducing training time compared with competing recurrent neural networks [Bai et al., 2018]. Dilated convolutional kernels [Chollet, 2017] are also applied to the design of a TCN to increase the receptive field in order to integrate global context. Without processing successive samples in a signal as a vanilla convolution, a dilated convolutional kernel touches the signal at every l^{th} entry.

Applying the time-domain speech separation system as a front-end for multi-talker speech recognition has been shown to effectively improve performance. Evaluated on WSJ0-2mix data,

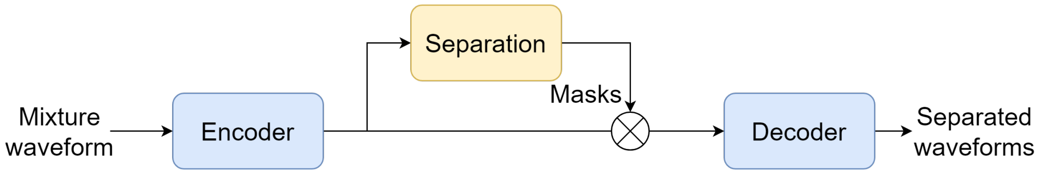


Figure 2.1: Diagram of end-to-end time-domain speech separation

using a RNN based end-to-end CTC/Attention speech recognition model to decode separated signals from a TasNet achieves 22.9% WER [von Neumann et al., 2020]. By jointly training the separation network with the acoustic model to optimise the speech recognition objective function, the WER is further reduced to 11.0% [von Neumann et al., 2020].

A thorough study has been conducted by Heitkaemper et al. [2020] to analyse the gain of each component presented in the TasNet compared with a common u-PIT frequency-domain separation approach. It has been shown that the short window length of the encoder, the SI-SNR objective function, and the learned encoder and decoder individually provide performance gains for the end-to-end approach in a simulated anechoic condition. However, in a more challenging condition with simulated reverberation, the fixed STFT and its inverse outperforms the learned kernels, highlighting the difficulty of training the encoder and decoder in a reverberant environment.

To further improve the modelling capacity of the separation network, advanced network architectures have been designed. The TCN has been replaced with U-Net blocks, which conducts consecutive downsampling and upsampling operations [Tzinis et al., 2020b]. A dual-path framework has been developed for modelling long-sequence signals [Luo et al., 2020]. It splits a long sequential input into multiple short chunks and iteratively processes intra- and inter- chunks. Recently, an attention mechanism has also been introduced into the separation system by using Transformer or Conformer blocks to construct the network [Chen et al., 2021; Subakan et al., 2021].

Most aforementioned single-channel separation methods perform well in anechoic environments without noise. However, their performance drops significantly when mixture signals contain noise and reverberation [Heitkaemper et al., 2020]. This limits the performance of the single-channel separation method when applied to distant microphone recordings. Nowadays, distant speech signals are usually recorded by using microphone arrays which are equipped with multiple microphones to record multiple sample-synchronised signals at the same time.

In the following section, we will review how multi-channel recordings from a microphone array have been exploited to improve speech separation performance in realistic environments with reverberation and noise.

2.3 Multi-channel speech separation

In the previous section, we have seen techniques for separating mixtures captured by a single microphone. In this section, we will see how the performance of separation approaches can be improved by exploiting multi-channel signals recorded by multiple devices. By using multiple recording devices positioned at multiple fixed locations, we can observe multiple 'views' of the same audio signal. Due to the way in which audio propagates with an acoustic space, the way that the sources are mixed in each channel is subtly different since the sources reach each channel by different routes and at different times. If the relative locations of the recording devices are known, these differences can be used to help separate sound sources that are located at different positions in the environment.

This section will review two contrasting approaches. First, we will look at classic signal processing solutions coming from the field of microphone array research and beamforming. We will then consider more recent approaches that use deep learning to exploit spatial cues directly from the mixture, typically with no explicit model of room acoustic being required.

2.3.1 Signal processing based approaches

One of the most effective multi-channel processing approaches for speech enhancement and separation is spatial filtering, which is also known as beamforming [Kellermann, 2008]. Beamforming combines the signals received from multiple devices into one enhanced single signal which is coming from at a particular angle while suppress signals at the other angles. The weighted delay-and-sum technique is one of the simplest multi-channel approaches operating on the time-domain signal [Anguera et al., 2007]. The weights and shifts for each microphone channel is based on time delay of arrival which is estimated based on a cross-correlation method. Signals received by different microphones are shifted and weighted by the estimated delays and weights and then summed together to focus on a desired direction. Adaptive beamforming approaches update the filter coefficients based on the estimated statistics of source and noise signals. Common adaptive beamforming approaches include generalized eigenvalue (GEV) [Warsitz and Haeb-Umbach, 2007], minimum variance distortionless response (MVDR) [Cox et al., 1987], generalized sidelobe canceler (GSC) [Griffiths and Jim, 1982], and linearly constrained minimum variance (LCMV) [Benesty et al., 2007]. Among them, the MVDR and GEV have been shown to yield promising results for speech processing [Barker et al., 2017].

Both the MVDR and GEV work in the frequency domain. Given an observation $\mathbf{Y}(k, f) = [\mathrm{Y}^1(k, f), \ldots, \mathrm{Y}^J(k, f)]^T$ received by a microphone array consisting of J microphones, a beamforming approach aims to estimate spatial filtering weights $\mathbf{w}$ to combine the multi-channel observation to obtain the target signal.

$$\hat{S}(k, f) = \mathbf{w}^H \mathbf{Y}(k, f) \tag{2.10}$$

The MVDR [Cox et al., 1987] aims to minimise the total energy of the output signal while at the same time preserves the desired signal in a specific direction. The design can be formulated as:

$$\min_{\mathbf{w}} \mathbf{w}^H \phi_{NN} \mathbf{w} \quad \textit{subject to} \quad \mathbf{w}^H \mathbf{d} = 1 \tag{2.11}$$

where $\mathbf{w}$ is the beamforming weight vector, ϕ_{NN} is the noise cross power spectral matrix, and $\mathbf{d}$ is the steering vector of the microphone array. $\mathbf{w}^H \mathbf{d} = 1$ preserves the desired signal in a specific direction. The MVDR solution of this optimal equation is:

$$\mathbf{w}_{MVDR} = \frac{\phi_{NN}^{-1} \mathbf{d}}{\mathbf{d}^H \phi_{NN}^{-1} \mathbf{d}} \tag{2.12}$$

A conventional method to estimate the steering vector uses the time difference of arrival (TDOA) between the microphones [Cauchi et al., 2015]. The ϕ_{NN} can be calculated from noise frames detected by using a voice activity detection (VAD) technique. However, the TDOA is difficult to estimate in a reverberant environment and as a result the estimated steering vector is imprecise. The MVDR beamformer is sensitive to the error in the steering vector estimation and the process will cause more deterioration than enhancement [Khabbazibasmenj et al., 2012]. In contrast, GEV maximises the output signal-to-noise ratio, without requirement of an accurate direction estimation [Warsitz and Haeb-Umbach, 2007]. The GEV solution can be expressed as:

$$\mathbf{w}_{GEV} = \operatorname*{argmax}_{\mathbf{w}} \frac{\mathbf{w}^H \phi_{SS} \mathbf{w}}{\mathbf{w}^H \phi_{NN} \mathbf{w}} \tag{2.13}$$

where ϕ_{SS} is the source signal cross power spectral matrix. The optimisation problem leads to a generalised eigenvalue problem, and the optimal beamformer is the generalised principle component. Although this method requires no estimation of acoustic transfer function between the source and receivers, the performance highly depends on the accuracy of the estimation of power spectral density matrices of both source and noise.

The source and noise cross power spectral matrices can be derived from spectral masks for speech and noise [Araki et al., 2004]. The masks can be estimated by clustering the frequency bins belonging to either sources or noise. Statistical models such as the GMM and its variants form a spatial mixture model to cluster each bin based on spatial features. The TDOA at each time-frequency bin is used to estimate binary masks in Araki et al. [2004]. Some other commonly used spatial features include unit norm, inter-channel phase difference (IPD) [Rickard, 2007], and inter-channel level difference (ILD). The IPD feature [Rickard, 2007] represents the phase difference between a pair of observed signals in the time-frequency domain, calculated as

$$\mathrm{IPD}^{(\mathrm{ij})}(\mathrm{k},\mathrm{f}) = \angle \mathrm{Y}^{\mathrm{i}}(\mathrm{k},\mathrm{f}) - \angle \mathrm{Y}^{\mathrm{j}}(\mathrm{k},\mathrm{f}), \tag{2.14}$$

Given the observations from a microphone array, the unit norm $\mathbf{Z}(k, f)$ is a directional statistical feature used in Sawada et al. [2011].

$$\mathbf{Z}(k, f) = \frac{\mathbf{Y}(k, f)}{\|\mathbf{Y}(k, f)\|} \tag{2.15}$$

Recently, the unit norm feature has been modelled by a complex angular Gaussian mixture model for speech separation task [Ito et al., 2016]. Following this work, a guided source separation system was developed as front-end processing for conversational speech recognition in real home environments [Boeddecker et al., 2018]. This method significantly reduced the WER from 81.1% to 62.5% on the CHiME-5 dev set. However, this guided source separation method has two main drawbacks. The first is that it requires a large context length, i.e. at least two seconds previous and after each target utterance, to achieve reliable separation performance. The second is that this model is a speaker-dependent model, which requires large amounts of speech signals from target speakers and cannot be applied to unseen speakers.

2.3.2 Deep learning based approaches

Deep neural networks have also been explored for the multi-channel enhancement and separation to take advantage of the more powerful modelling capability in comparison to the Gaussian mixture models [Heymann et al., 2016]. Heymann et al. [2016] proposed an approach that employed a neural network to estimate speech and noise spectral masks for each single-channel. The obtained masks were combined via a median filter to compute the cross-power spectral matrices for computing beamforming coefficients. This approach has been successfully applied to the CHiME-3 task, outperforming a conventional delay-and-sum beamformer and a MVDR beamformer without employing neural networks. However, the neural network based mask estimator in this algorithm only used spectral information, while the spatial information that provides complement location information was ignored.

One advantage of a neural network model is that it can automatically fuse information from multiple modalities to learn a representation for a target task. Taking advantages of this power, neural networks have been explored to exploit the spatial information in addition to the spectral information to estimate either masks or learn direct spectral mapping for speech separation [Wang and Wang, 2018]. Wang et al. [2018] extracted IPD features from multi-channel signals and included the spatial features into a deep clustering model [Wang et al., 2018]. In simulated reverberant conditions, the multi-channel deep clustering yielded a higher SDR score than a single-channel deep clustering system and conventional spatial clustering approaches, such as MESSL [Mandel et al., 2010] and GCC-NMF [Wood et al., 2017]. Alternatively, PIT is effectively extended to multi-microphone input by taking IPDs features as additional input [Yoshioka et al., 2018a]. A network is used to estimate masks for each source, after which beamforming weights are calculated based on the masks and applied to the multi-channel signal for better separation performance. The beamformer helps reduce processing artifacts that degrade speech recognition performance, in comparison to direct multiplying the estimated masks on the mixture signal.

Most frequency-domain approaches only focus on estimating magnitude related masks and ignore the enhancement of phase information. Although beamforming techniques can recover the phase partially from the mixture signal, inaccurate phase information degrades the quality of the reconstructed signals. However, it has more recently been found that accurate phase estimation is useful for better spatial processing such as beamforming [Wang et al., 2020]. To enhance both magnitude and phase, a direct complex spectral mapping approach is studied in Wang et al. [2021]. The complex representation of the input is split into the real and imaginary parts, and the separation model is trained to estimate the real and imaginary parts of the source.

Inspired by the success of end-to-end approaches to single-channel time-domain speech separation as described in Section 2.2.3, multi-channel separation has also explored the end-to-end framework. A straightforward way is to extend the single-channel time-domain framework to the multiple microphone by inputting the time-domain signals from all the available channels to the network [Gu et al., 2019]. Specifically, the single-channel encoder is repeated to process each channel to generate encoded representations. Manually extracted spatial features have also been incorporated into the TasNet to provide spatial information [Gu et al., 2019]. To be more specific, the IPD features on the frequency domain are upsampled and concatenated to the spectral information encoded from the single-channel signal.

Combination of the time-domain separation network and beamforming has also been investigated in Ochiai et al. [2020]. The separated signals from a single-channel time-domain separation network are transformed into the time-frequency domain to compute the power spectral matrix for beamforming. Integrating the beamforming process into the network design has also been studied, in which the beamforming coefficients are learned from neural networks [Luo et al., 2019; Zhang et al., 2021c]. A filter-and-sum network (FaSNet) has been proposed to learn time-domain adaptive beamforming filters to combine all microphone channels [Luo et al., 2019]. This approach is to approximate the conventional delay-and-sum operation [Anguera et al., 2007]. An all deep learning MVDR beamformer is proposed to use a recurrent neural network to replace the matrix inversion and principle component analysis involved in the MVDR [Zhang et al., 2021c].

Although combining a time-domain single-channel separation network with conventional spatial features can benefit the separation performance [Gu et al., 2019], using the IPD features in a single end-to-end time-domain approach may cause mismatch and misalignment problems. This is because the window length of the convolutional kernel used for extracting spectral features is much smaller than the STFT window length used for extracting IPD features. It has been demonstrated in Wölfel and Woelfel [2009] that such feature concatenation suffers in performance if features from different streams are not perfectly synchronised.

Besides, there is lack of study about whether spatial features can be effectively extracted from time-domain multi-channel signals by using neural networks. This calls for a novel fully neural network design for end-to-end time-domain speech separation using multiple recordings. Considering that the STFT and inverse STFT can be replaced with trainable kernels in the time-domain single-channel separation system, questions arise as to whether spatial features can be extracted in a similar fashion via the neural network based encoder, and whether the learned features can benefit speech separation performance.

For end-to-end single-channel separation system, various advanced network architectures have been designed to improve separation performance [Luo et al., 2020; Tzinis et al., 2020b]. However, they have not been successfully applied to multi-channel systems yet. The question of how to exploit advanced network structures to better capture spatial and spectral information for speech separation is worthy of further investigation.

2.4 Speech extraction with speaker information

Most source separation techniques aim to segregate all sources given an observation. The design of such a source separation system typically assumes there exists prior knowledge of the number of sources in the observation. However, in practice, not all sources are of interest to us, especially the noise and late reverberation of speech. Furthermore, it is more difficult for a separation system to reconstruct ambient noise signals from a mixture than speech signals. An alternative approach, speaker extraction, exploits speaker identity information to select one or a few target sources to extract from a mixture. A speaker extraction system has three major advantages over a source separation system: 1) The speaker identity of the separated signal is controlled and known in the extraction system. 2) There is no label permutation problem in the extraction system. 3) There is no requirement for a speaker extraction system to know the number of speakers in the mixtures in advance. This section will describe progress that has been recently made on deep learning based speaker extraction systems.

Conventionally, speech extraction systems are built based on a speaker-dependent model. A speaker-dependent extraction system is trained to output one specific speaker from a mixture signal [Du et al., 2014]. To extract different speakers, multiple models should be developed independently, each of which corresponds to one speaker. Building such a speaker-dependent system requires a large number of training samples of mixture signals consisting of a fixed target speaker and various interfering speakers. Furthermore, the speaker-dependent model cannot be applied to situations where target speakers are unseen in the training set.

Recently, progress has been made in the design of extraction framework that can operate in a speaker-independent fashion [Wang et al., 2019b; Žmolíková et al., 2019; Žmolíková et al., 2017]. In this case, the extraction system is informed explicitly by the identity information of a target speaker to be extracted from a mixture. Crucially, the target speaker does not have to be present in the training data. This model can generalise well to the extraction of unseen speakers on the condition that the target speaker information is available and provided at the time of extraction. This new framework usually consists of a speaker embedding system and a source separation system. The source separation system is biased with the information about the identity extracted from the embedding system to identify and trace the target speaker in a mixture signal. This behaviour is similar to how humans attend to one object based on the characteristics of this target [Brungart, 2005].

The first speaker-independent extraction system was presented in Žmolíková et al. [2017]. This extraction system employed a neural network for target speaker mask estimation, which was then used for computing beamfoming coefficients. The extraction system was informed by auxiliary speaker information, which was extracted from an adaptation utterance generated from the target speaker. It was shown that the extraction system can separate speech signals belonging to one specific speaker from a multi-speaker mixture. This multi-channel extraction system has also been shown to improve ASR performance in a multi-talker reverberant scenario [Žmolíková et al., 2018]. Following this work, the informed speaker extraction scheme has been demonstrated to be efficient for single-channel speaker extraction as well [Delcroix et al., 2018].

The speaker identity information is usually represented as an embedding vector with a fixed dimension, which is generated by using the embedding network to process an enrollment non-overlapped signal uttered by the target speaker. The embedding network can be either jointly trained with the separation system to minimise the signal reconstruction loss [Žmolíková et al., 2019] or trained in a different task, i.e., a speaker recognition task [Wang et al., 2019b; Xu et al., 2019b]. I-vector is a commonly used speaker representation for speaker verification [Dehak et al., 2011] and has been applied to a speaker extraction system for target signal reconstruction [Xu et al., 2019a]. Wang et al. [2019b] used an embedding network that is specifically designed for end-to-end speaker recognition. The embedding network is trained with a large-scale data designed for speaker recognition. Using separately trained embedding network provides an opportunity to make use of additional data to increase speaker variations, which can potentially make the speaker embedding more discriminative. However, it was argued that since the embedding network is separately trained for a different task, the obtained speaker embedding may not capture the most appropriate speaker information for the propose of target speaker enhancement [Ji et al., 2020].

How to provide an extraction system with the speaker embedding vector still remains an open question. To inform the extraction system with the identity information, the speaker embedding can be inserted to the extraction network via factorised layer [Delcroix et al., 2018], scaling adaptation [Delcroix et al., 2019], concatenation operations [Xu et al., 2019a], or an attention mechanism [Ochiai et al., 2019]. The factorised layer and scaling adaptation approaches are similar to a gate function that multiplies embedding components on the internal representation in the extraction network. However, the concatenation approach repeatedly concatenates the embedding vector with the input features in an extraction system. The neural network is inspired to learn speaker-dependent transforms from the concatenated features. Ochiai et al. [2019] unified separation and extraction systems and used an attention mechanism to select target speaker components from internal representations in the separation system.

Most aforementioned extraction systems operate on time-frequency domain. Motivated by the success achieved by the end-to-end time-domain separation approach, time-domain extraction systems have been developed recently [Delcroix et al., 2020; Xu et al., 2020a, 2019b]. In a time-domain extraction system, the STFT and its inverse operations are replaced with trainable layers, e.g. one dimensional convolutional layers. The time-domain extraction system is trained to directly optimise a signal construction objective on the time-domain such as SI-SNR. It has been shown that the time-domain approach significantly improve the quality of the extracted signal in terms of SDR, compared to conventional T-F domain approaches [Delcroix et al., 2020].

Instead of enhancing only one speaker each time, some studies attempt to extract multiple speakers at the same time. To extract all participating speakers, a speaker inventory, i.e. a list of pre-enrolled signals of candidate speakers, has been used to inform a separation system [Wang et al., 2019a]. Since the speaker inventory contains both target and non-target speakers, the correlation scores between the estimated mixture embedding and the speaker embedding in the inventory are calculated and used to select the relevant speakers. A speaker-aware separation system is developed in Xu et al. [2020b]. A speaker information encoder is used to disentangle speaker identity information from the mixture, which is used to inform a mask inference network. A speaker-conditional chain model (SCCM) has been proposed that firstly infers speaker identities, then uses the corresponding speaker embeddings to extract all sources [Shi et al., 2020]. A similar approach, Wavesplit, learns embedding from a mixture signal using a clustering method to inform a subsequent separation system [Zeghidour and Grangier, 2021]. However, they are trained with the PIT criterion without explicitly conditioning the output order on the embeddings. Therefore, speaker identities of the separated signals in the multi-speaker extraction systems are still arbitrary.

Lastly, recordings from multiple microphones have been exploited to improve extraction systems. Žmolíková et al. [2017] exploited the GEV beamformer for the multi-channel speaker extraction. Additional spatial features extracted from a microphone array have also been shown to improve the extraction systems in both clean and reverberant environments [Delcroix et al., 2020]. Delcroix et al. [2020] incorporated IPDs features as additional input to a time-domain extraction model to better discriminate speakers in reverberant environments. The IPDs are inserted into the middle-stage features after the speaker identity features in the extraction system.

The current end-to-end multi-channel extraction systems still rely on conventional IPD features, which are not optimal for the end-to-end framework. Recently, spatial features directly learned from neural networks were investigated to improve time-domain speaker extraction systems [Zorila et al., 2021]. Specifically, a 2-D convolutional layer was used to construct a spatial encoder to capture spatial information from the time-domain multi-channel input. Although this study has shown that the spatial features can improve extraction systems, the multi-channel extraction systems did not provide further gain over multi-channel separation systems. This points out that the current design of a multi-channel extraction system cannot fully exploit spatial information and speaker identity information. It is hypothesised in this thesis that the degradation comes from improper fusion strategies for features from multiple modalities.

Using deep neural network in data fusion provides a large flexibility, allowing multiple modalities combined at different stages, namely an early stage, an intermediate stage, or a late stage. Although features from different modalities can be fused simultaneously in a single layer, this simple scheme may not be able to learn the relationship between each modality. It was shown in Neverova et al. [2016] that highly correlated modalities should be fused first, after which less correlated ones are progressively fused. The current design of end-to-end time-domain speaker extraction system has not fully considered the fusion strategy for spectral information, spatial information, and speaker identity information. This question leads to the research conducted in Chapter 5.

2.5 Unsupervised and semi-supervised learning for speech separation

The separation and extraction methods mentioned above are developed with supervised learning and they work well when evaluated with matched simulated data. However, when they are evaluated with real mixtures, the performances are usually poor due to the mismatches between the real and simulated environments. This is because a deep neural network based system is sensitive to the mismatch. This cannot be fixed simply by retraining with 'real data' as in real scenarios there is generally no access to isolated ground truth signals. Collecting such data is expensive and requires recordings with controlled conditions.

Although a speech separation model can be jointly trained with a speech recognition system to optimise speech recognition objectives when the main goal is to improve speech recognition performance, transcribing large amounts of noisy speech takes a lot of human resources. Further, training a combined separation and recognition network from scratch may not lead to the best recognition performance. It has been shown that a good initialisation for each module followed by joint fine-tuning outperforms training all modules together with random initialisation [Chen et al., 2018a]. Therefore, developing speech separation systems in real scenarios is of great practical importance for distant speech recognition and other speech applications.

A major problem with supervised approaches is that the simulated mixtures generated during training, fail to anticipate the variety of signals observed during testing. Some methods attempt to enlarge the training set with enough varieties of acoustic conditions [Maciejewski et al., 2019]. However, the infinite environment conditions in the real world cannot be easily covered by a finite training set. Maciejewski et al. [2019] shows that although multi-condition training with enlarged dataset greatly improves performance in near-field conditions, significant gaps remain in far-field conditions.

Instead of relying on strong labels like isolated sources, weak labels have been exploited for training an audio sound separation system [Pishdadian et al., 2020a,b]. The used weak label can be a sound activity label, which indicates time periods where a sound is active or not. This approach has been developed for the audio sound separation task, in which different types of sound have distinguished sound characteristics. However, such a weak label may not be enough to train a speech separation system, because the difference between each speaker is less discriminative than the difference between each sound class, such as car horn and dog bark. Further, this approach still relies on annotated labels and cannot be applied to situations where audio mixtures are without any labels.

To exploit unlabelled data, unsupervised learning approaches have been explored. One approach is to exploit multiple modalities to generate weak labels for training a separation network. When speech signals are recorded by multiple microphones, unsupervised spatial clustering can be used as the first step to predict labels for training a neural network based separation system [Bando et al., 2019; Drude et al., 2019a; Tzinis et al., 2019]. Drude et al. [2019a] used a complex angular central Gaussian mixture model (cACGMM) to group time-frequency representations into clusters to obtain class labels, which are then used to train a deep clustering model. This work was later extended to a joint optimisation of the cACGMM and a neural network mask estimator [Drude et al., 2019c]. In Tzinis et al. [2019], spatial features, i.e. IPDs, were clustered via K-means to generate class labels for training a deep clustering model. The trained deep clustering model is able to perform separation on a single-channel recording.

The labels predicted from an unsupervised spatial clustering approach are shown to be effective to train a classification based separation model, i.e. deep clustering [Drude et al., 2019a]. However, the predicted labels may not be good enough to guide a regression based speech separation model, because the current unsupervised spatial clustering method cannot totally remove residual interference speech and noise components in separated signals. Since a regression based separation model optimises signal reconstruction, using a noisy target during training confuses the separation model. The performance of a regression based separation network trained with noisy single speech as target still falls far behind models trained with clean ground truth [Maciejewski et al., 2021].

Since recent state-of-the-art separation performance is achieved by the end-to-end time-domain separation system [Luo and Mesgarani, 2019], efforts have been made to develop unsupervised approaches following this trend. Instead of relying on weak labels or multiple modalities for supervision, an approach, named as mixture invariant training (MixIT), attempts to directly train an end-to-end separation network on noisy signals [Wisdom et al., 2020]. This approach uses the mixture signals to generate more noisy signals and trains a separation model to reconstruct the original mixture signals from the corrupted one [Wisdom et al., 2020]. The MixIT has achieved similar speech separation performances compared to a fully supervised system in both anechoic and reverberant conditions. However, this technique suffers an 'over-separation' problem caused by a training data mismatch: the training data - a mixture of mixtures - will contain more sources than the individual mixtures that it is being trained to segregate. The artificially corrupted input signal fails to match the SNR distribution to the real test signal again and this mismatch has a negative impact on the separation performance.

An alternative approach to address domain mismatch is semi-supervised learning. Lam et al. [2020] proposed a Mixup-Breakdown training framework that used a pre-trained supervised model to process noisy signals from a mismatched domain to generate pseudo references. Then the prediction can be used to train a student model by exploiting consistency constraints. Specifically, the same noisy signal can be augmented with various strategies and the student model is trained to provide consistent prediction with the teacher model. In addition, the teacher model parameters are updated as an exponential moving average of the student model parameters to generate more accurate predictions [Tarvainen and Valpola, 2017]. This semi-supervised separation system has been shown effective to separate speech when the noise type in the target domain is unseen in the source domain. However, this method is only evaluated with different types of noise which is simply added to the speech signals through simulation. It is unclear if the teacher model can still provide reliable prediction when the total environmental conditions change between source and target domains.

Most current unsupervised approaches designed for end-to-end time-domain separation systems still suffer mismatch between training and testing. Fon instance, the SI-SNR conditions for training and testing data for the MixIT approach are significantly different. For the Mixup-

Breakdown training, the teacher model is still trained from simulated data, and it may not be able to provide reliable references from real mixtures when the target domain is severely mismatched to the source domain.

Another practical issue is that the existing unsupervised and semi-supervised learning frameworks have not fully considered realistic conditions with noise and reverberation. The MixIT approach was evaluated on reverberant speech separation and the Mixup-Breakdown training was evaluated on noisy speech without reverberation. In far field conditions, both reverberation and noise should be considered together and this will pose a more challenging task compared to considering each factor individually. For supervised models, multi-channel recordings have been demonstrated to be effective to improve separation performance in far-field conditions. However, there is lack of study about how to train an end-to-end multi-channel separation network in situations where supervised training data is limited. Therefore, Chapter 6 will investigate two questions: 1) How to alleviate the mismatch problem in the current state-of-the-art unsupervised learning approaches for speech separation? 2) How to develop unsupervised and semi-supervised learning approaches for end-to-end multi-channel separation models?

2.6 Summary

This chapter has provided a thorough overview of progress made in distant speech recognition in everyday environments in recent years and the development of speech separation techniques. Important aspects in the existing body of literature and the leading research questions are summarised as follows.

- Many efforts have been made at improving speech recognition performance in noisy and reverberant conditions. However, the problem caused by overlapped speech has been ignored in most cases due to there being no overlap in mainstream read speech corpora and little overlap in heavily studied meeting scenarios. In everyday environments where conversations are conducted by multiple speakers, the overlap ratio could increase and dramatically degrade the recognition performance. To understand better the impact of overlapped speech on the conversational speech recognition, Chapter 3 will conduct a thorough analysis using a recently released multi-talker party corpus, namely the CHiME-

5.

- The end-to-end separation framework has yielded promising results compared to traditional signal processing approaches and deep learning approaches operating on the frequency-domain. However, there are remaining questions concerning how to build an end-to-end multi-channel separation system. It is unknown how to use neural networks to encode spatial features from time-domain signals for speech separation. And it is unclear if advanced structure developed for single-channel separation could benefit the multi-channel separation as well.

- In contrast to blind speech separation, speaker extraction recovers particular speakers based on additional speaker identity information. However, existing speaker extraction systems suffers from improper strategies of feature fusion from multiple modalities. This issue specifically limits the performance of time-domain multi-channel speaker extraction systems. Therefore, Chapter 5 will investigate strategies concerning how to exploit speaker information in a time-domain multi-channel speaker extraction system.

- Unsupervised and semi-supervised learning strategies have been proposed to build deep learning based speech separation systems without access to clean reference sources. In particular, the mixture invariant training framework can train a single-channel separation model in a fully unsupervised way and achieve a competitive result compared with a supervised learning trained model. However, existing unsupervised and semi-supervised approaches have not fully addressed mismatch problems between training and testing. And they are not extended to the end-to-end multi-channel separation system yet. These two questions lead the research conducted in Chapter 6.

In total, the study of these above-mentioned research questions will help expand the knowledge of the areas of speech separation and overlapped speech recognition. Individually, the first study will provide a better understanding concerning overlapped speech in the natural conversations in daily life. The analysis will reveal the difficulty of conversational speech recognition and motivate the later research. The other studies concerning speech separation techniques will not only advance the speech separation area, especially the design of multi-channel separation system and the application to real scenarios, but also lead to a better usage of speech separation techniques as a front-end for the distant speech recognition in everyday environments.

Chapter 3

The influence of overlapping speech on automatic speech recognition

3.1 Introduction

Automatic speech recognition in reverberant and noisy conditions has made great progress in recent years. This can be attributed in part to advances in speech processing and audio enhancement, but also to the availability of real speech corpora recorded in challenging environments [Barker et al., 2017; Vincent et al., 2017]. However, when multiple speakers talk at the same time, recognising speech that is overlapped by competing speakers still remains a challenging task.

Early studies concerning multi-talker speech recognition in real scenarios mainly use natural conversations recorded in meeting settings [Renals and Swietojanski, 2017; Swietojanski et al., 2013; Yoshioka et al., 2018b]. For example, using the AMI corpus of multiparty meeting, the recognition performances of individual headset microphones and multiple distant microphones are 25.6% and 49.4% [Renals and Swietojanski, 2017]. In the distant microphone case, the WER for the non-overlapped segments is around 40%, while a absolute 8-12% increase is observed when considering segments with overlapped speech.

Recently, it has been observed that the performance of the state-of-the-art ASR systems that are designed for single-talker speech drop significantly when evaluated with spontaneous conversations in real home scenarios [Barker et al., 2018]. This has been clearly demonstrated in the CHiME-5 challenge [Barker et al., 2018], which provides the first large-scale corpus of multi-talker conversational speech recorded in real home environments. The CHiME-5 dataset contains near-field and distant microphone recordings. The recognition performances of the near-field data and distant microphone data with a state-of-the-art hybrid DNN-HMM acoustic model are 47.9% and 81.3%, respectively [Barker et al., 2018]. These recognition results are much worse than the performance achieved with meeting conversations and fully demonstrate the difficulty of recognising spontaneous conversational speech. It was observed that speech overlaps happen very frequently in the CHiME-5 recordings and might pose a challenge to the speech recogniser. This calls for a thorough investigation of the extent of overlapping speech in casual conversations and how it impacts on ASR system performance.

With the aim of improving speech recognition performance in casual conversational settings like CHiME-5, this chapter will analyse the performance of two contrasting front-end approaches, namely channel selection and speech separation. Channel selection approaches work by selecting individual signals from amongst multiple recording devices. Speech separation approaches try to separate speech sources from a noisy mixture.

Channel selection strategies can be very effective in the distributed microphone or distributed microphone array settings that are becoming more commonly used to record speech signals in a large space such as meeting rooms or home environments. The quality of a target speech recorded by each microphone varies, due to the distance between the source and each receiver, resulting in different degrees of distortions caused by noise and reverberation. To exploit distributed microphones to improve speech recognition performance, channel selection methods aim to select the signal that leads to lowest WER. The selection can be based on signal quality measurement [Guerrero et al., 2018; Wolf and Nadeu, 2014], or information related to the ASR system [Wölfel, 2007; Xiong et al., 2018]. However, although channel selection strategies are conceptually simple, they can be hard to apply in practice. In particular, in real home settings, speakers move frequently and distance between speech sources and microphones is large, which make it difficult to select signal of good quality from available devices.

An alternative approach is to use a source separation system as a pre-processor for an ASR system. In multi-talker meeting scenarios, source separation techniques have been developed to address problems caused by overlapping speech [Araki et al., 2016; Yoshioka et al., 2018b]. In home environments, speech separation could be more challenging compared to meeting scenarios, due to more frequent speaker movements, larger distances between source and receivers, and the greater variety of interfering noise types. Source separation systems have been shown to be valuable as front-ends for CHiME-5 ASR systems [Boeddecker et al., 2018]. One of the best performing approaches is a signal processing based speech separation method [Boeddecker et al., 2018]. This separation approach employs a complex angular central Gaussian distribution based spatial mixture model to estimate time-frequency mask for a target speaker, which is used to compute the coefficients of an MVDR beamformer. This chapter will analyse the performance of this source separation approach and channel selection strategies with respect to the difficulties posed by overlapping speech.

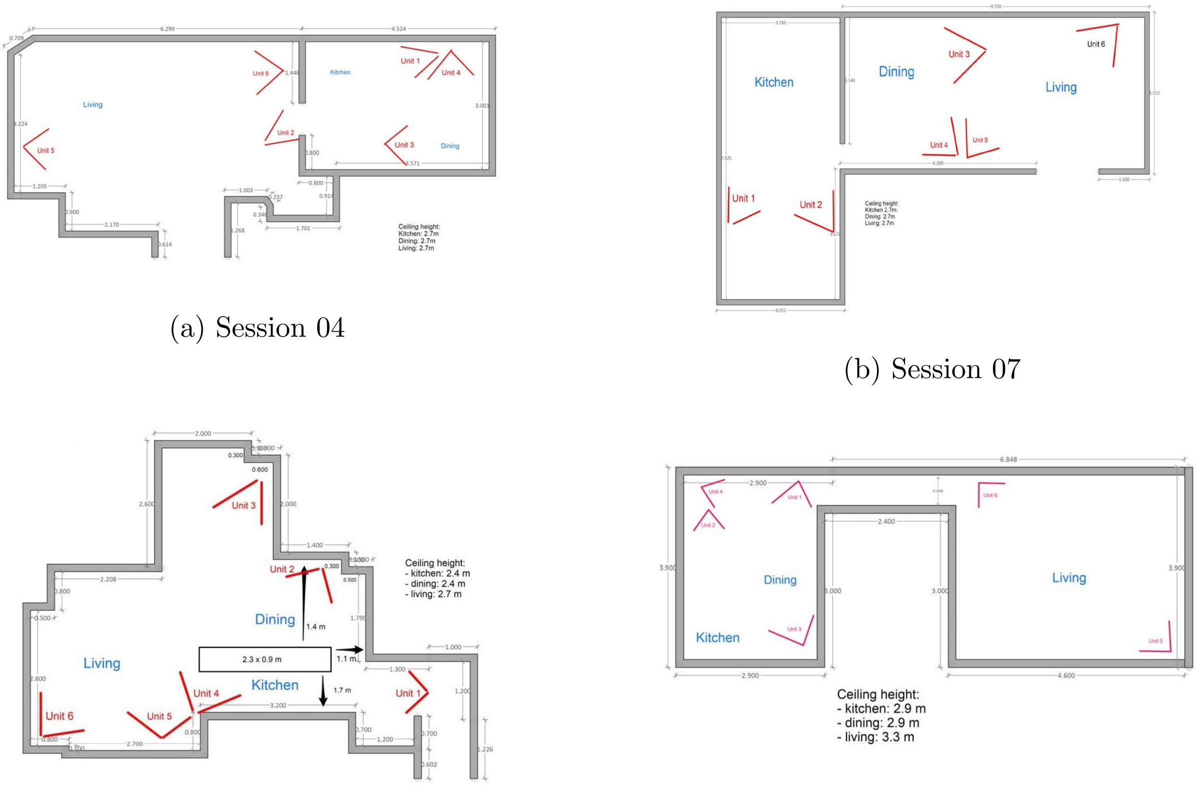
This chapter first presents an overview of the CHiME-5 data which records natural conversations from multiple talkers in dinner-party scenarios. Section 3.3 then analyses the extent of overlapping speech in natural conversations. Section 3.4 conducts analysis concerning how the overlapping speech impacts on the recognition performance in casual conversations. Section 3.5 investigates whether a channel selection strategy in distributed microphone settings can sufficiently address the problem caused by overlapping speech. Finally, Section 3.6 extensively studies an existing signal processing based speech separation method for speech recognition in real home settings.

3.2 Overview of the CHiME-5 dataset

The CHiME-5 dataset, designed for ASR evaluation, provides the first large-scale corpus of natural conversational speech among multiple speakers in real home scenarios. Full details of the dataset can be found in Barker et al. [2018], but are summarised here for convenience. The data has been solicited using a ‘dinner party’ scenario. Each dinner party has four participants, two acting as hosts and two as guests. Crucially, the party members are all friends who know each other well and the parties are taking place in the participants’ own homes. The familiarity of the situation means that the communication and behaviours have been made as natural as possible. There are 20 parties in total, each of which represents a unique acoustic environment and lasts two to three hours. Each party consists of three separated stages, each lasting at least 30 minutes. The three stages contain a *kitchen* phase for food preparation, a *dining* phase for eating food, and a *living* phase for post-dinner socialising. Each stage typically corresponds to a different location within the living space (i.e., kitchen, dinning area, living room). Participants have been allowed to move naturally from one location to another during the party and converse on any topic they choose with no constraints. After collection, the transcripts and audio recordings are redacted to remove personally identifying utterances and to remove material likely to be deemed offensive. Participants were given the opportunity to review the audio and transcripts as part of the consenting process.

The speech material is recorded by six distant Kinect microphone arrays placed at different locations and four in-ear binaural microphones worn by the participants. Each distant microphone array is a linear array of four sample-synchronized microphones. The Kinect devices send signals to independent laptop computers. This approach allowed the recording system to be set up rapidly within people's homes but means that the signals between arrays have not been well synchronized during recording. The six devices have been placed in a way that there are at least two devices capturing the activity in each stage. The worn device is a set of Soundman OKM II Classic Studio binaural microphones, which is mainly used to facilitate time annotations and utterance transcription. The dataset is distributed with roughly sketched floorplans. Figure 3.1 shows examples of the floorplans in four different recording sessions. The locations of the distant microphone arrays, labeled as Units, are illustrated in the floorplans. Note that the red lines indicate the direction and field of view of the Kinect devices. Video was also recorded to aid with transcription but only the audio signals have been made publicly available.

The 20 sessions are separated into three parts, namely, a training dataset, a development dataset and an evaluation dataset containing 16 sessions, 2 sessions and 2 sessions, respectively. The size of each dataset is given in Table 3.1. In the training sessions, there are two parties for each group of four speakers, but each party is in a different location. Therefore, the total number of speakers for the 16 parties in the training set is 32. The audio recorded by both worn microphones and distant microphone arrays in the training dataset can be used for building the speech recognition system. For the development set and evaluation set, an official reference array from the distant microphone arrays has been selected for each evaluation utterance. The reference array depends on the session and locations and is chosen to be in the same room as the speakers. However, the reference device is not necessarily the best device, or even the closest device. That is because a lot of the living spaces are open-plan (e.g., in Figure 3.1. (a)



(a) Session 04

(b) Session 07

(c) Session 12

(d) Session 23

Figure 3.1: Floorplan examples in CHiME-5, showing the position and direction of the microphone arrays, the areas for kitchen, dinning and living. All dimensions are in meters. The floor plans are not all the the same scale.

and (c) there is a kitchen-dining area whereas in (c) the living and dining area are connected).

For each utterance, a reference transcription saved in a JSON format is provided, in which time annotations including start time, end time, current speaker, and the word sequence have been manually obtained by professional transcribers listening to the speaker's binaural recording. For each other recording device, the start of the recordings were synchronised using an audible 'beep' played in the room. This is not precise because of the different time delay of arrival of the beep to the various devices but should synchronise signals within about 5 ms. The signals then lost synchronisation over the 2-3 hour recording due to clock drift and frame dropping. The clock drift happens because the clock in each Kinect did not run at exactly the same rate, which caused +-50-100 ms shift over 2-3 hours. The frame dropping could be 2-3 seconds in the worst cases. To compensate for the synchrony across devices, an array synchronisation post-processing has been performed prior to the publication of the data, which corrects

Table 3.1: Size of each dataset in the CHiME-5 data

Dataset	**Parties**	**Speakers**	**Hours**	**Utterances**
Train	16	32	40:33	79,980
Dev	2	8	4:27	7,440
Eval	2	8	5:12	11,028

the start and end time of the utterance by shifting the reference timings by an amount that has been estimated using a cross-correlation approach [Barker et al., 2018].

This corpus captures a number of acoustic factors that exist in real life, which are challenging to an ASR system. The difficulties can be categorised into two factors. The first factor is the reverberation and background noise that were introduced to the recordings because the Kinect microphone arrays recorded speech signals at a distance from the source speakers. From Figure 3.1 showing the floor plans, it can be inferred that even the closest Kinect devices are often two or three meters from the speakers. In addition, the speakers will often not be directed towards the closest Kinect. The second factor is the casual conversational speech style, which is typified by frequent stretches of hypo-articulated speech, highly colloquial language, and a high degree of speech overlap. Speaker overlaps from recorded meetings have proven to degrade an ASR system because of both acoustic cross-talk and different speech styles between overlap speech and non-overlap speech [Cetin and Shriberg, 2006]. Compared with meeting scenarios, natural spontaneous conversations contain more casual speech and speaker change, potentially resulting in more speaker overlaps.

3.3 Overlapping speech analysis

In this section, we aim to analyse the degree of speech overlap in multi-talker casual conversations. The analysis will first concern the duration within a whole conversation where various numbers of speakers are simultaneously active. Then, the analysis is performed at the utterance level, statistically estimating the duration of speaker overlap in each utterance and the proportion of overlapped speech in each utterance. The CHiME-5 data described in the previous section is used for the analysis.

As mentioned in Section 3.2, the CHiME-5 data is recorded in 20 parties, each of which

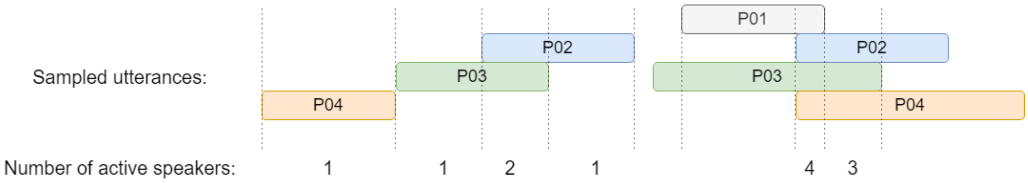


Figure 3.2: speech segments extraction visualisation.

lasts two to three hours. We first determine in each party the periods containing non-speech, single-talker speech, and overlapped speech. The time stamps (the start-time and the end-time) of all utterances provided in JSON transcriptions are first sorted in each party. Then, the whole party conversation is split into segments based on every two consecutive time stamps. In each segment, we count the number of active speakers based on the speaker activity provided in the JSON transcription. Since each party has four participants, the number of active speakers ranges from zero to four. Figure 3.2 illustrates an example of segmented conversations covering several utterances in one session. The number of active speakers are indicated at the below of each segment. For each utterance, the overlapping duration $t_{overlap}$ is calculated by summing the periods that have more than one active speaker within the utterance. The overlap proportion r of each utterance is defined as the ratio of overlapping duration $t_{overlap}$ to the whole duration t_{utt} of the reference utterance.

$$r = \frac{t_{overlap}}{t_{utt}} \tag{3.1}$$

The first analysis concerns the duration in each party where multiple speakers are simultaneously active. Figure 3.3 illustrates the N-speaker active duration for a representative sample of 9 of the 20 parties. The single-speaker active state is the most common state. However, it can be observed that events with multiple active speakers happen in every dinner party. The degree of overlap between participants varies among different parties. Specifically, in Session 13 and Session 16, there is only one active speaker for most of the time. In comparison, the duration of more than one active speakers in Session 02 and Session 08 is much longer than other parties, which could be caused by heated discussions or multiple conversations happening simultaneously.

Table 3.2 shows the proportion of non-speech and speech with various numbers of speakers in each individual session. The distribution of these proportions are calculated across all 20 sessions and shown in Figure 3.4. It can be observed that single-talker speech dominates most of the time during a party, occupying, on average, 55% of each party's literal duration. Over-

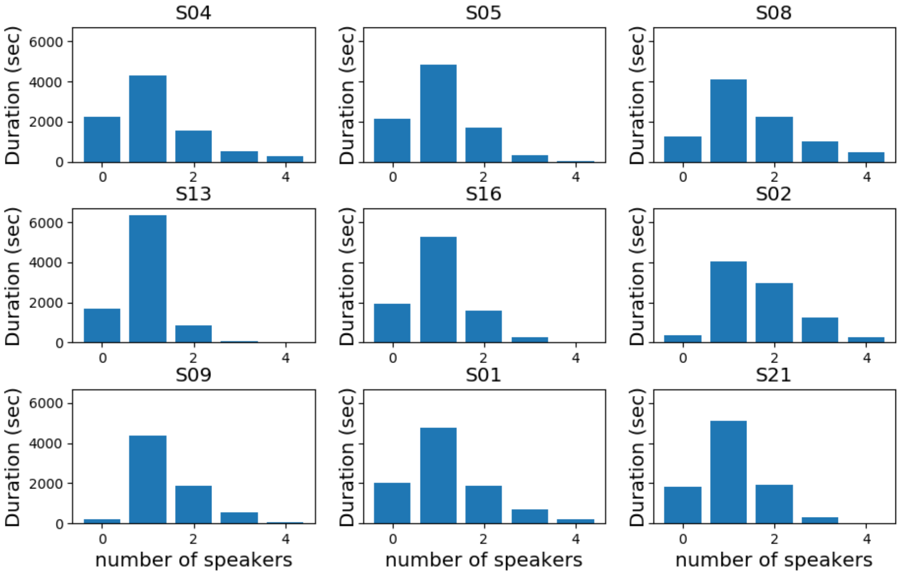


Figure 3.3: The total duration in each party where either 0, 1, 2, 3, or 4 speakers are simultaneously active

lapping speech is occurring around 20% of the time during a party, during which a foreground speaker is overlapped by speech from one or more talkers. Therefore, for nearly 30% of the time during which speech is occurring, one or more competing talkers are also active, which could be expected to have a significant impact on ASR performance if not explicitly dealt with. This proportion can even reach as high as 40%, for example in Session 02. This proportion in this natural conversation situation is higher than previous reported figures that have been estimated using data in both multiparty meetings and telephone conversations. For example, Cetin and Shriberg [2006] reported on data in 26 different meetings from NIST that around 12% of all foreground speaking time could be overlapped by one or more competing talkers. Even though each party involves four participants, most of the overlaps involve only one background speaker.

To better understand how overlap speech affects each utterance, the duration of overlapped speech in each utterance is estimated. Figure 3.5 shows the distribution of duration of overlapped speech in each utterance in sampled parties. In most cases, the overlap duration for each utterance is shorter than 0.5 seconds. It is observed that the longest duration of overlapped speech is usually shorter than four seconds. In Session 02 and Session 09, most of the overlap duration is longer than 0.5 seconds, demonstrating that there is more overlap speech in these two parties than others.

Table 3.2: Proportion of non-speech, single-talker speech, and simultaneous talk with 2 speakers and more than two speakers. Superscripts † and $*$ indicate sessions belonging to the development set and the evaluation set respectively.

Session	**0 speakers**	**1 speaker**	**Overlap**	
			2 speakers	**More than 2**
S01*	0.21	0.50	0.19	0.10
S02†	0.04	0.46	0.33	0.17
S03	0.31	0.52	0.15	0.02
S04	0.25	0.48	0.18	0.09
S05	0.24	0.54	0.19	0.04
S06	0.18	0.59	0.20	0.03
S07	0.18	0.65	0.16	0.01
S08	0.14	0.46	0.24	0.16
S09†	0.03	0.61	0.27	0.09
S12	0.25	0.66	0.08	0.01
S13	0.19	0.71	0.09	0.01
S16	0.22	0.58	0.17	0.03
S17	0.17	0.53	0.22	0.08
S18	0.09	0.63	0.23	0.05
S19	0.21	0.55	0.20	0.04
S20	0.16	0.44	0.23	0.17
S21*	0.20	0.56	0.21	0.03
S22	0.16	0.46	0.24	0.14
S23	0.23	0.49	0.22	0.06
S24	0.30	0.51	0.17	0.02

The duration of overlap speech is used to calculate the proportion of overlap speech in each utterance. In fact, this short duration could be caused by extended time margins during transcribing rather than the true speaking time. Note, transcribers tend to err on the side of extending the speech segment in this way as they want to make sure that the annotated segments do not clip the start or the end of the utterance. To reduce the effect caused by possibly extended margins in the time annotation, the boundary of each utterance is adjusted by delaying the start time by 200 ms and setting the end time 200 ms earlier. This adjustment can also reduce the influence caused by backchannel speech, which is usually shorter than 0.5 seconds and is fully overlapped by longer utterances from other speakers.

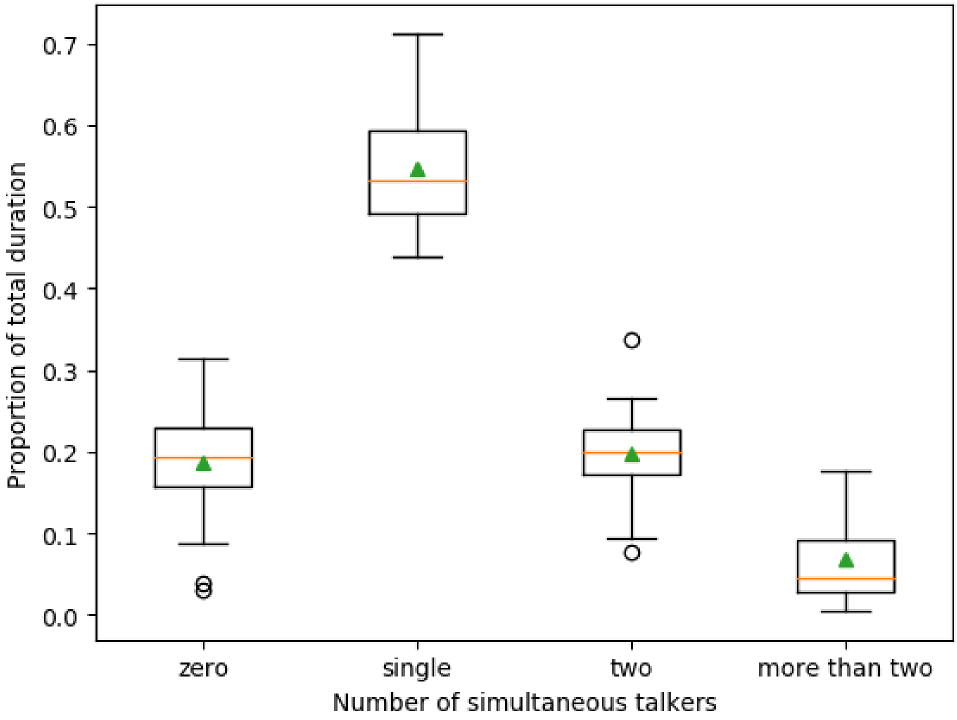


Figure 3.4: Rates of non-speech, single-talker speech, two-talker speech, and speech from more than two talkers. The green triangles show the mean of each distribution and the orange line shows the median.

Figure 3.6 shows the distribution of the overlap proportion in each utterance. Among the utterances with overlapped speech, the proportion of overlap within one utterance distributes approximately uniformly. In most of parties, a high percentage of utterances are not overlapped by competing speakers. However, in Session 02 and Session 09, the number of utterances with overlapped speech is much larger than the number of utterances without overlapped speech. That is to say, most of utterances during the whole party are overlapped by competing speakers. By listening to the audio recordings, we observed that two separate conversations happened at the same time in Session 02 (but in the same space, i.e. all four voices are captured by the same recording devices) and participants in Session 09 were having heated discussions. To study the effect of overlapped speech on the ASR performance, these two sessions will be used for ASR evaluation. The correlation between the recognition accuracy and the proportion of overlapped speech will be shown in the next section.

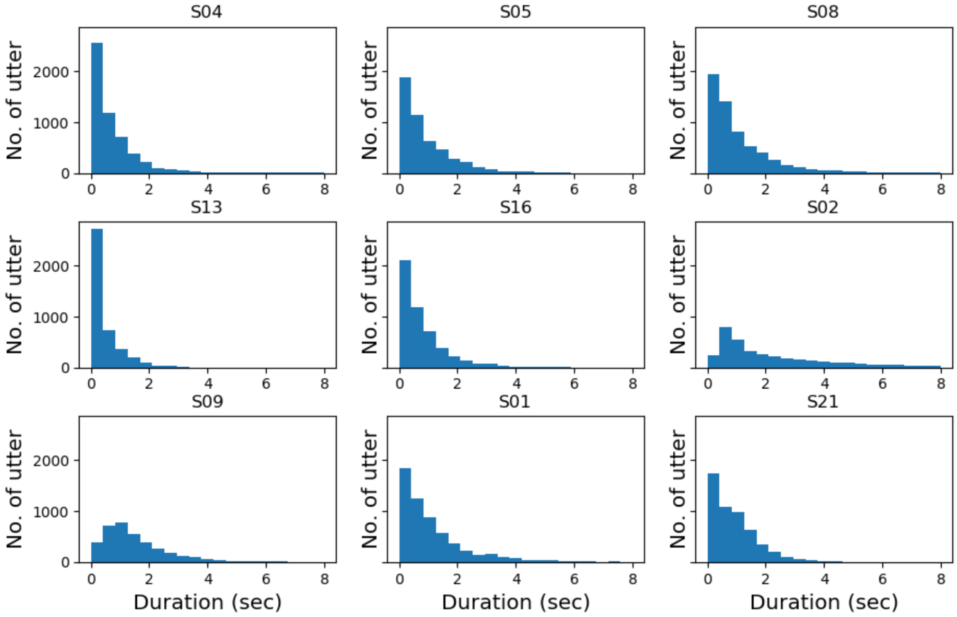


Figure 3.5: The duration of overlapped speech in each utterance

3.4 Impact of overlapped speech on ASR

In this section, we aim to analyse the impact of overlapped speech on speech recognition performance in real home environments. The CHiME-5 data described in Section 3.2 is selected to evaluate a state-of-the-art speech recognition system. The following parts will first describe the ASR system. Then, we will study how the performance is impacted by overlapped speech.

3.4.1 ASR system

To evaluate the speech recognition performance, an ASR system has been employed in this work, which consists of an acoustic model, a language model, and a speech enhancement system. The acoustic model uses a 15-layered Factorised TDNN (TDNN-F) structure [Povey et al., 2018], where each TDNN-F block has a layer with 1536 units and a bottleneck layer with 160 units. This structure is selected here since the same acoustic model design has achieved high performance on a conversational telephone speech corpus, i.e. Switchboard, yielding 12.4% of WER [Povey et al., 2018]. The input features to this model are Mel-Frequency cepstral coefficients (MFCC) extracted from audio signals and a 100-dimensional i-Vector. The acoustic model is trained with a Lattice-free Maximum Mutual Information (LF-MMI) [Povey et al.,

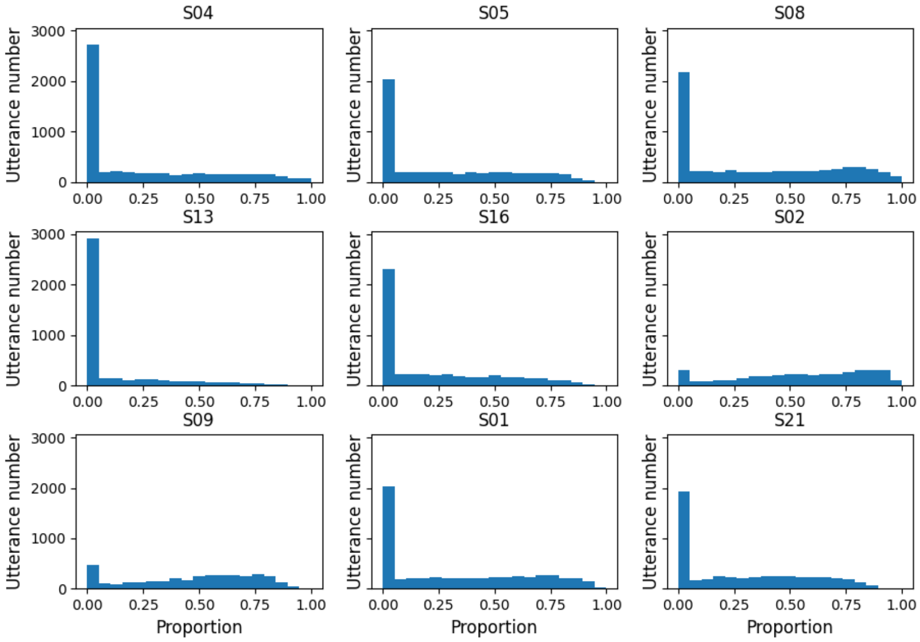


Figure 3.6: Proportion of overlapped speech in each utterance.

2016] objective function to predict target HMM states, which are obtained using a pre-trained GMM-HMM system to produce a forced alignment. The test sets are decoded in two stages for robust i-Vector extraction [Manohar et al., 2019]. The implementation of the acoustic model is based on Kaldi [Povey et al., 2011]. The training set is composed of the worn microphone recording from the left channel of the binaural microphones and a subset of 100k randomly-selected utterances from the distant microphone array data. The training data is augmented by speed perturbation [Ko et al., 2015] with ratio 0.9 and 1.1.

The language model provided in the CHiME-5 challenge is used. A CMU dictionary[1] is used as a basic pronunciation dictionary. Since the CHiME-5 conversations are spontaneous speech, a number of words are not present in the CMU dictionary. To provide the pronunciations of these out-of-vocabulary words, grapheme-to-phoneme conversion is used based on Phonetisaurus G2P [Novak et al., 2012][2]. The language model is automatically selected based on perplexity on training data, and a 3-gram language model is selected, which is trained by the maximum entropy modelling method as implemented in the SRILM toolkit [Alumäe and Kurimo, 2010; Stolcke, 2002; Wu and Khudanpur, 2002].

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[2] https://github.com/AdolfVonKleist/Phonetisaurus

To reduce the effect caused by background noise and reverberation, speech enhancement methods are applied to the distant recordings in the test data. The speech enhancement methods include the weighted prediction error (WPE) [Nakatani et al., 2010] and a delay-and-sum beamformer [Anguera et al., 2007]. The implementation of these two methods is based on the released baseline recipe for the CHiME-6 challenge [Watanabe et al., 2020]. The multiple input multiple output version of the WPE method is used for dereverberation [Nakatani et al., 2010]. WPE estimates the reverberation tail and subtracts it from the observed signal. The dereverberation method is performed by using the reference array that contains four microphones. The reference array information is provided in the JSON transcription file. The weighted delay-and-sum beamformer is performed by using the dereverberated multi-channel signals to output a single-channel signal. The enhanced distant signal is subsequently fed into the acoustic model for decoding. For the in-ear microphone, the raw signal recorded by the left ear is used for decoding.

3.4.2 Results and analysis

The speech recognition performance of the TDNN-F acoustic model is presented in Table 3.3. The table shows the overall performance and the separate performances achieved in each of three sub-phases (Kitchen, Dining, Living) of the CHiME-5 development sessions. The ASR performance for in-ear microphone signal and distant array recordings are 43.03% and 74.14% of WER, respectively. The large performance gap is caused by the different degree of speech distortion observed on the two types of recordings. The speech signal received by the binaural microphone worn by the speaker can be considered to be approximately similar to a clean speech signal. Since the worn microphone is close to the target speaker, there is little reverberation and the background noise recorded by the microphone and the speech from interfering speakers has little impact on the speech from the target speaker because of the relatively high signal to noise ratio. For the signals received by a distant microphone array, the target speech can be masked by both ambient noise and interference speech, and the reverberation will also degrade the quality of the received speech signal. These factors are harmful to a speech recognition system, making the result obtained with the distant microphones much worse than that from close-talking microphone.

Table 3.3: WER (%) for the baseline system with Factorised TDNN structure.

Model	Session	Kitchen	Dining	Living	Overall
Worn	S02	46.33	48.35	41.27	43.03
	S09	40.84	43.20	37.82	
Kinect	S02	82.35	73.28	69.89	74.14
	S09	73.78	74.05	70.56	

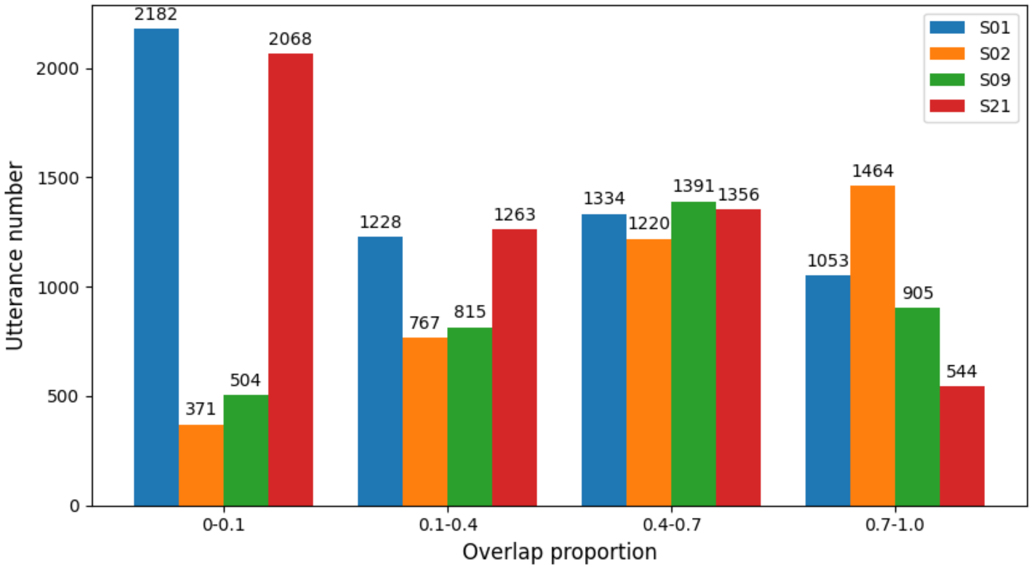


Figure 3.7: Number of utterances of different conversational conditions based on overlap proportion for the development (S02 and S09) and evaluation (S01 and S21) sets of CHiME-5

To better understand the effect of overlapped speech on the recognition performance, the test data is split into different conversational conditions specified by the proportion of overlapped speech in each utterance. Four conditions are considered: a non-overlap condition with overlap proportion (r) smaller than 0.1, a mild overlap condition (0.1<r<0.4), a medium overlap condition (0.4<r<0.7), and a severe condition where r is greater than 0.7. Figure 3.7 shows the number of utterances in each condition for the development set and evaluation set.

Speech recognition performance has been measured for each of these test sets. Table 3.4 shows the detailed results with both the worn and the Kinect set-ups. We first observe that in the non-overlap speech condition, there is a performance gap between the close-talking microphone and the distant microphone. This gap is possibly due to the background noise and reverberation in the environments. The second observation is that the recognition performance drops for both close-talking and distant microphones when the overlapped speech proportion

Table 3.4: WER(%) under various conditions of overlapped speech proportion in each utterance

	Non-overlap	**Mild**	**Medium**	**Severe**
	(r<0.1)	**(0.1<r<0.4)**	**(0.4<r<0.7)**	**(r>0.7)**
Session 2				
Worn	36.3	38.9	44.2	51.0
Kinect	60.3	62.8	74.3	88.7
Session 9				
Worn	32.6	33.3	45.0	53.9
Kinect	55.2	58.9	82.1	103.4

increases. When the overlap proportion increases, the performance gap also increases and reaches its peak under the severe condition where most of utterances are fully overlapped by interfering speech. This observation suggests that the overlapped speech has a larger impact on recognising distant recordings than recognising close-talking microphone signals.

3.5 The effect of channel selection

In the previous section we have seen overlapped speech severely degrades performance of distant speech recognition. To look for solutions, this section investigates the extent to which channel selection methods can reduce the impact of overlapped speech. The analysis will employ clean reference signals to select a distant signal with the best speech intelligibility among all available devices, because a signal of higher speech intelligibility is assumed to lead to better recognition performance. Although clean reference signals are not accessible in practice, this selection strategy is being used here to provide an upper bound estimate on the best ASR performance a channel selection strategy can achieve, therefore showing whether selection methods along are sufficient to address the problem of speech overlap in home environments. As an additional 'oracle' experiment, we also consider ASR decoding based selection, i.e. directly using the transcriptions and choosing the channel that has the lowest word error rate, which will provide an even more optimistic upper bound performance for channel selection. Blind channel separation approaches are not covered in this section, but readers who are interested in this topic can refer to Xiong et al. [2018].

3.5.1 Speech intelligibility based channel selection

We first use the clean speech signal as a reference for channel selection. A speech intelligibility measure based on correlation between clean and degraded speech, namely the Short-Time Objective Intelligibility (STOI) [Taal et al., 2011], is used to select the best channel from all available distant arrays. The STOI score is selected because it was shown to be highly correlated with the speech intelligibility of noisy speech. Also, the STOI score only exploits the magnitude of time-frequency domain signals to calculate the correlation. The STOI measurement is more suitable than direct signal domain cross-correlation between noisy speech and clean speech because the large distance between microphones weakens the correlation among signals especially for phase information.

The STOI score is calculated with a 10 kHz sampling rate, therefore signals of other samplerate should be resampled first. Then the resampled audio signal is transformed into time-frequency domain using a 512-point STFT. A Hann window with a size of 256 samples and a hop of 128 samples is used when computing the STFT. The STFT representations are transformed into an octave band representation using 15 one-third octave bands that extend up to approximately 4,300 Hz. Since silent regions do not contribute to speech intelligibility, they are detected by using the clean signal and excluded from both clean and noisy signals. In the clean speech, when the energy of a frame is lower than 40 dB compared to the maximum frame energy, this frame is determined as a silent frame.

Let $\hat{\mathbf{X}}_{j,m}$ and $\hat{\mathbf{Y}}$ denote the energy in the j-th one-third octave band at the m-th time frame of degraded speech and clean speech, respectively. The intermediate STOI measure for one bin, $d_{j,m}$, depends on a neighborhood of N previous bins. To do so, new vectors $\mathbf{X}_{j,m}$ and $\mathbf{Y}_{j,m}$ are constructed, which consist of $N = 30$ previous frames before the m-th time frame as follows.

$$\mathbf{X}_{j,m} = [\hat{\mathbf{X}}_{j,m-N+1}, \hat{\mathbf{X}}_{j,m-N+2}, \ldots, \hat{\mathbf{X}}_{j,m}]^T \tag{3.2}$$

Then, $\mathbf{X}_{j,m}$ is scaled and clipped as follows.

$$\bar{\mathbf{X}}_{j,m} = \min(\frac{\|\mathbf{Y}_{j,m}\|}{\|\mathbf{X}_{j,m}\|}\mathbf{X}_{j,m}(n), (1 + 10^{-\beta/20})\mathbf{Y}_{j,m}(n)) \tag{3.3}$$

where $\mathbf{X}_{j,m}(n)$ denotes the n-th value in $\mathbf{X}_{j,m}$. The β refers to the lower signal-to-distortion bound and is set to 15 dB. The intermediate intelligibility matrix $d_{j,m}$ is computed as the correlation between $\bar{\mathbf{X}}_{j,m}$ and $\mathbf{Y}_{j,m}$.

$$d_{j,m} = \frac{(\bar{\mathbf{X}}_{j,m} - \mu_{\bar{\mathbf{X}}_{j,m}})^T(\mathbf{Y}_{j,m} - \mu_{\mathbf{Y}_{j,m}})}{\|\bar{\mathbf{X}}_{j,m} - \mu_{\bar{\mathbf{X}}_{j,m}}\| \cdot \|\mathbf{Y}_{j,m} - \mu_{\mathbf{Y}_{j,m}}\|} \tag{3.4}$$

where $\mu_{(\cdot)}$ is the sample average of the corresponding vector. The final STOI score is obtained by averaging the intermediate intelligibility measure over all time frames and all bands:

$$d = \frac{1}{JM}\sum_{jm} d_{j,m} \tag{3.5}$$

Based on the STOI score, we aim to select one best channel from six distant arrays for each utterance in the CHiME-5 development set. Before calculating the STOI score, the WPE and BeamformIt are applied to each distant array, resulting in six candidate channels for selection. The left channel of the binaural signal is used as the clean speech. Then a STOI score is computed between each processed distant signal and the clean speech signal. The processed distant signal with the highest STOI score is selected, presumed that it is likely to yield the best recognition score.

We also perform channel selection based on the ASR decoding by accessing transcriptions. The same enhancement method is applied to all distant array signals. Then the six enhanced signals are decoded and the one achieving the lowest WER is selected.

3.5.2 Experiments and results

Speech recognition performance with the channel selection methods are shown in Table 3.5. It shows that the official reference array provided in the JSON transcription file is not the best channel for speech recognition of active speakers. The signal selected based on the STOI score yields around 4.2% WER reduction, which emphasises the importance of channel selection in this distributed microphone/arrays scenario. WER based selection yields almost 10% WER reduction compared with STOI based selection. This result can be seen as an upper bound performance for the channel selection approach. However, when one utterance is very short with only one or two words, selecting a channel that has only recorded noise signals is likely to have fewer errors compared with a channel recording both target speech and interfering speech. In practice, the performance of WER based selection is hard to achieve with the signal metric based selection.

Table 3.5: WER (%) for the TDNN-F system with channel selection

Model	**Session**	**Kitchen**	**Dining**	**Living**	**Overall**
Baseline	S02	82.35	73.28	69.89	74.14
	S09	73.78	74.05	70.56	
STOI	S02	71.61	69.53	65.79	69.97
	S09	72.92	72.42	70.90	
Oracle	S02	64.29	61.47	57.59	60.41
	S09	58.97	60.56	60.08	

Table 3.6: WER(%) under various conditions of overlapped speech proportion in each utterance

	Non-overlap	**Mild**	**Medium**	**Severe**
	(r<0.1)	**(0.1<r<0.4)**	**(0.4<r<0.7)**	**(r>0.7)**
Session 2				
Baseline	60.3	62.8	74.3	88.7
STOI	53.6	58.1	67.9	81.3
Session 9				
Baseline	55.2	58.9	82.1	103.4
STOI	52.9	57.1	81.8	106.4

Table 3.6 presents the channel selection performance in different conditions of overlapped speech. With the STOI based channel selection, the ASR performance for overlapping speech is still far behind non-overlapping speech performance. In Session 09, the channel selection strategy shows limited benefits to the overlapping speech recognition, compared to using fixed microphone array as shown in the baseline. It is possibly due to that when all microphone devices are positioned at a distance to speakers, even if the best channel has been selected, interfering speech still exists in the signal. This suggests that selection strategies alone cannot fully solve the overlapping speech problem in real home settings.

3.6 The effect of source separation

An alternative solution to reducing the impact of overlapped speech is speech separation, which has been applied to multiparty meeting recordings to reduce the impact of speech overlap [Araki et al., 2016; Yoshioka et al., 2018b]. This section will analyse the effectiveness of a state-of-

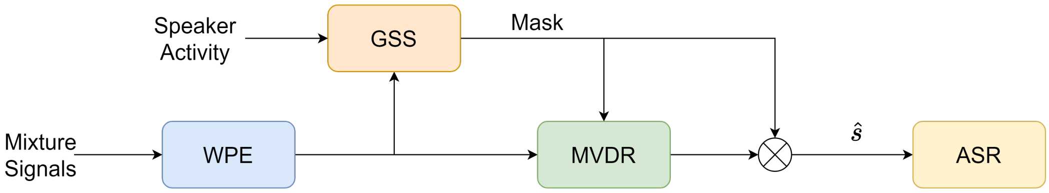


Figure 3.8: Overview of the speech enhancement system consisting of dereverberation and source separation

the-art separation technique for dealing with the problem caused by overlapped speech. An existing speech separation method, namely guided source separation (GSS) has been applied to the speech recognition task using CHiME-5 data in Boeddecker et al. [2018]. This part provides a more detailed analysis that concerns how separation techniques perform under various conditions of overlapped speech, and results from the separation approach will be compared with results from the channel selection approach.

3.6.1 Guided Source Separation

The whole speech enhancement system consists of three stages, a dereverberation stage, a source separation stage, and a beamforming stage, as illustrated in Figure 3.8. In the first stage, a multi-channel WPE [Yoshioka and Nakatani, 2012] is applied to the microphone signals for dereverberation. The second stage is the GSS [Boeddecker et al., 2018], which exploits both spatial information extracted from microphone arrays and speaker identity information. For the GSS, a complex angular central Gaussian distributions [Ito et al., 2016] based spatial mixture model aims to extract the source speech from the dereverberated mixture signal. The GSS uses prior knowledge of speaker activity to initialise the spatial mixture model. An expectation maximisation (EM) algorithm estimates the parameters of the spatial mixture model. The outputs of the spatial mixture model represents the posterior probabilities of each speaker being active. In the third stage, the obtained posterior probabilities are used for calculating coefficients of a minimum variance distortionless response (MVDR) beamformer [Souden et al., 2010]. The reference channel for the beamformer is estimated based on an SNR criterion [Erdogan et al., 2016], which will select the best device when multiple microphone arrays are available. The extracted target speech is later fed into an acoustic model for decoding.

When applying the GSS to the CHiME-5 dataset for evaluation, we use the same configuration described in Kanda et al. [2019]. It was shown that, for good EM initialisation, it is crucial to have access to a long window of acoustic context preceding and following the utterance [Boeddecker et al., 2018], therefore, 15 seconds previous and after each considered utterance is used for WPE and GSS. For the spatial covariance matrix estimation in the MVDR beamformer, only the original utterance length is considered, since using a large context may cause problems when speakers are moving [Kanda et al., 2019]. Both single-array and multi-array enhancements have been performed. For the single-array enhancement, the method is performed by using the reference array in the JSON transcription file. All four channels of the array are used as input. For the multi-array enhancement, the microphone signals from all arrays are stacked to form one big array that contains 24 channels. The JSON transcription file provides the time annotations that indicate when a particular speaker is active, which are used for both initialising the EM algorithm and solving the frequency permutation and the global speaker permutation problem.

3.6.2 Experiments and results

Table 3.7 shows the ASR performance when GSS is applied. For the single-array system, the source separation technique effectively reduces the WER compared to the baseline system. It shows greater benefit to the recognition system when the target speech is overlapped by competing speakers in comparison to the non-overlap condition, meaning that the effect of overlapped speech on the recognition system is reduced. When accessing all available channels, the GSS can further improve the performance, and the gain becomes larger when the degree of overlap is more severe. This is because increasing the number of available channels can provide more spatial information for the GSS to distinguish between sources.

Compared to the channel selection approach, the speech separation approach is more effective when dealing with overlapping speech. GSS using multiple arrays provides a much larger gain of WER in all conditions than the STOI based channel selection approach. This is because the multi-array separation method can not only select the microphone device that contains speech of good quality, but also separate the speech belonging to the target speaker out of interfering speech and noise. This observation demonstrates that a strong speech separation technique is necessary for performing distant speech recognition in real home environments.

Table 3.7: Recognition performance with GSS as front-end processing under various conditions of overlapped speech proportion in each utterance

	Non-overlap (r<0.1)	**Mild (0.1<r<0.4)**	**Medium (0.4<r<0.7)**	**Severe (r>0.7)**
Session 2				
Baseline	60.3	62.8	74.3	88.7
STOI	53.6	58.1	67.9	81.3
GSS (Single-array)	54.2	56.0	66.6	81.1
GSS (Multi-array)	47.2	48.6	53.1	63.8
Worn	36.3	38.9	44.2	51.0
Session 9				
Baseline	55.2	58.9	82.1	103.4
STOI	52.9	57.1	81.8	106.4
GSS (Single-array)	51.9	53.8	75.3	98.0
GSS (Multi-array)	45.1	45.5	61.1	80.6
Worn	32.6	33.3	45.0	53.9

Although GSS has achieved significant progress for overlapped speech recognition, there are some limitations. The gap of performance between the close-talking microphone signal and the far-field signal still remains large and the overall performance by using distant microphone arrays (around 50% WER) cannot meet the practical requirements. The recognition performance with overlapped speech is still much worse than the performance of non-overlapped speech. This issue calls for the design of stronger separation techniques, which should also be robust to the background noise and reverberation in real environments. Another issue is that GSS requires a large temporal context of 15 seconds left and right of the considered segment to achieve reliable separation performance [Boeddecker et al., 2018]. However, having such a large context may become problematic when the speakers are moving and is not acceptable for real-time applications.

3.7 Summary

In this chapter, the extent of overlapped speech in a natural conversational scenario has been analysed with natural conversations recorded in multi-talker dinner scenarios. Analysis shows that speaker overlaps happen frequently and take around 30% of time of conversational speech. It is also shown that the overlapped speech significantly degrades the performance of a speech recognition system. When the proportion of overlapped speech in an utterance increases, the ASR accuracy drops.

A channel selection method and a separation method have been investigated for reducing the effect caused by overlapped speech. A speech intelligibility based selection method by accessing clean reference signals achieves 4% WER reduction compared to using fixed microphone arrays. This result shows that the gain from the channel selection approach is limited in real home environments. While the speech separation approach using multiple devices effectively enhanced the target speech in situations where competing speakers exist. Evaluation using the CHiME-5 data shows that the separation method provides 12-20% WER reduction over the channel selection when dealing with the overlapped speech.

However, it is observed that there is still a large performance gap between non-overlapped and severely overlapped speech with the signal processing based separation approach. Furthermore, this separation approach is non-casual and needs a large context length, around 15 seconds before and after one considered segment, to achieve good performance. These observations motivate us to design stronger speech separation techniques to overcome the difficulty caused by overlapped speech in the conversational speech recognition task.

Deep learning based speech separation techniques have achieved large progress and have been applied to overlapping speech recognition [Yoshioka et al., 2018a]. Recently, end-to-end time-domain speech separation models have achieved state-of-the-art separation performances over conventional signal processing based methods [Luo and Mesgarani, 2019]. Motivated by this, in the following chapters, we will begin to develop deep learning based time-domain multi-channel separation systems and investigate how the developed techniques perform for overlapping speech recognition.

It should also be noted that, in the following chapters, evaluation is conducted on simulated data that was described in Section 2.2.1, instead of using the CHiME-5. This is partially because, as mentioned in Section 2.2.1, simulated data allows access to oracle clean signals to build supervised speech separation systems and makes it convenient to analyse the performance. Another reason comes from the complexity and difficulty of the CHiME-5 data. Although the following chapters will address several research questions concerning deep learning based speech separation techniques and make large progress on simulated data by developing several strong separation systems, it is still difficult to directly apply these developed methods to real scenarios such as the CHiME-5. There are some other research questions remaining to be addressed, and the author suggests in the Chapter 7 to leave them as future work.

Chapter 4

Blind time-domain multi-channel speech separation

4.1 Introduction

The previous chapter has shown that overlapped speech occurs frequently in natural conversations and degrades the performance of automatic speech recognition systems. Using speech separation techniques as pre-processing can reduce the impact of overlapped speech. It has been shown that the guided source separation enhancement provides an almost absolute 20% WER reduction for the conversational speech recognition task with CHiME-5 data [Boeddecker et al., 2018]. However, the improved performance of overlapped speech is still much worse than the performance of non-overlapped speech. Therefore, better speech separation methods are needed.

Recent advances in deep learning have facilitated the development of very powerful source separation systems [Chen et al., 2017; Hershey et al., 2016; Isik et al., 2016; Kolbæk et al., 2017]. It is shown in Chapter 2 that the end-to-end single-channel speech separation approach that directly optimises time-domain signals has shown promising results and surpassed ideal binary masking methods [Luo and Mesgarani, 2019]. This method replaces the STFT operations with trainable modules, namely an adaptive encoder and decoder, which are jointly optimised with a separation module. It reduces the reconstruction errors caused by inaccurately estimated phase in the conventional time-frequency approaches.

End-to-end separation systems are relatively new approaches and there are many questions that remain open about their design. These questions include considerations on how to extend them to multi-channel scenarios, how to combine them with approaches to reverberation, how to best capture the long term temporal dependencies in the mixed signals, and how to reduce the artificial distortions caused by the separation process. These are among the questions that will be addressed in this chapter.

When multi-channel recordings are available, spatial information can be combined with spectral information as input features to benefit a separation network. Hand-crafted spatial features (e.g., inter-channel phase difference, IPD) are common spatial features used as additional input and have been shown to provide performance improvements for T-F domain based separation systems [Wang et al., 2018]. However, IPD features are not optimal for inclusion in an end-to-end system as they are extracted from the frequency domain and this leads to a domain mismatch problem. A recent work alleviates the cross-domain problem by including the IPD extraction as a learnable block built into the end-to-end system [Gu et al., 2019]. However, in this work, the window length used for learning the IPD extraction is different from the window length of the spectral encoder in the end-to-end system, which could cause a misalignment problem between the two types of feature.

In real environments, source separation and speech recognition are made more challenging by the effects of reverberation (i.e., acoustic reflections). The end-to-end approach has been shown to be effective in a single-channel dereverberation problem [Luo and Mesgarani, 2018a]. Joint separation and dereverberation has been performed by training a system to separate anechoic speech from a reverberant mixture [Delfarah and Wang, 2018, 2019]. However, training deep learning based dereverberation systems requires matched anechoic signals to act as training targets and the matched anechoic signals are difficult to collect in practice. In addition, these methods have not considered the impact of reverberation on the performance of separation systems.

End-to-end speech separation systems require a network capable of modelling extremely long input sequences. For a two second long audio segment with an 8 kHz sampling rate, there are 16,000 samples in the signal. Long signals are difficult to model for conventional sequential modelling networks such as RNNs because they cannot be easily trained when the sequence is long [Pascanu et al., 2013]. Many efforts have been made to improve neural networks' sequence modelling capacity for speech separation. The dual-path RNN splits the sequence of input features into short chunks and interleaves two RNNs for local and global modelling [Luo et al., 2020]. Some studies explored convolutional neural network (CNN) architectures for sequence modelling [Luo and Mesgarani, 2019; Tzinis et al., 2020b]. One predicted unit from a CNN layer relies on a region of the input and the region in the input is called the receptive field for that unit [Luo et al., 2016]. It is crucial to have a large receptive field size and this has been realised by either stacking multiple dilated 1-dimensional (1-D) convolutional blocks [Luo and Mesgarani, 2019] or by iteratively performing downsampling and upsampling operations [Tzinis et al., 2020b]. However, the problem of how to increase sequence modelling capacity for a *multi-channel* separation system remains relatively unexplored.

Signals separated from a deep-learning based separation system usually contain many artificial distortions and separation errors such as frames flipping. A speech recogniser is sensitive to unseen distortions, leading to recognition performance degradation. It was shown in Isik et al. [2016] that an enhancement post-processing step can effectively reduce the distortion introduced by a time-frequency separation system. The enhancement stage takes one of the separated signals and the original mixture to estimate the target speech again. The enhanced signals also yield better speech recognition performance than the signals without enhancement. Although time-domain separation approaches directly optimise the signal construction loss, the quality of the separated signals may still contain distortions and could be refined.

This chapter studies the end-to-end approach for multi-channel speech separation. The system is built upon the successful single-channel separation model, Conv-TasNet [Luo and Mesgarani, 2019]. Instead of using conventional IPDs as additional spatial features, a novel spatial encoder constructed by a 2-D convolutional layer is proposed here to learn more suitable spatial features for the end-to-end multi-channel system. In addition, three types of modification are considered to improve the proposed system in noisy and reverberant environments. The first modification concerns the effects of reverberation on source separation and speech recognition. In fact, it is observed during experiments that the estimation of spectral features and spatial features of the separation system declines in quality in reverberant environments. To allow better features to be extracted by the proposed multi-channel system, a dereverberation step is employed into the separation processing. The study investigates the best way of combining the time-domain separation and an existing dereverberation method, i.e., the weighted prediction error (WPE) [Nakatani et al., 2010]. The second modification investigates an alternative network architecture for the separation module within the multi-channel separation system. The iterative downsampling and upsampling operations [Tzinis et al., 2020b] are introduced to improve the sequence modelling capability and boost the performance. The last modification studies a multi-stage separation framework to reduce the distortions caused by separation processing. The separated signals will be enhanced by a time-domain enhancement network.

This chapter is organised as follows. Section 4.2 describes an existing multi-channel separation system that includes IPDs in the end-to-end separation system. Section 4.3 considers the design of an end-to-end approach for multi-channel separation and its extensions. Section 4.4 presents implementation details and describes the experimental setup. Section 4.5 presents the results and analysis. Section 4.6 summarises the major findings in this chapter.

4.2 IPDs based multi-channel separation

This section describes an existing approach that has been designed for extending an end-to-end separation model to form a multi-channel version, which will be used as a baseline in the chapter. This system, denoted as IPD-TasNet, integrates both the end-to-end separation network and IPD features [Gu et al., 2019]. The input to the separation system is an observation $\mathbf{y} \in \mathbb{R}^{J \times T}$ received by a microphone array containing J microphones. This system is illustrated in the Figure 4.1.

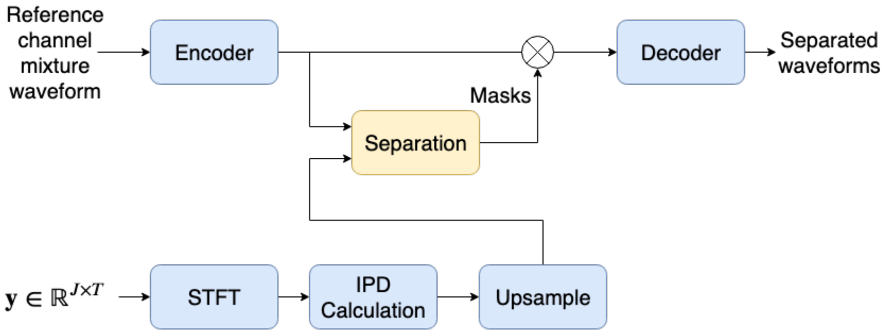


Figure 4.1: IPD based multi-channel speech separation approach.

Similar to the single-channel time-domain separation system, the IPD-TasNet uses a trainable encoder to encode a raw time-domain mixture from a reference microphone channel. Since the encoded representations mainly contains temporal-spectral information, this encoder can be called a *spectral encoder*. The spectral encoder is a 1-D convolutional layer with a kernel size of $(1, L)$, followed with a non-linear activation function, i.e., a rectified linear unit (ReLU). It transforms each segment of the mixture time-domain signal into an N-dimensional representation, where N is the output dimension of the spectral encoder.

The well-established IPD features are used to provide the system with spatial information. The IPD feature is computed as the phase difference between a pair of signals in the time-frequency domain:

$$\mathrm{IPD}_{\mathrm{nk}}^{(\mathrm{ij})} = \angle \mathrm{Y}_{\mathrm{nk}}^{\mathrm{i}} - \angle \mathrm{Y}_{\mathrm{nk}}^{\mathrm{j}}, \tag{4.1}$$

where Y_{nk}^{i} denotes the STFT of the i-th microphone signal at frame n and frequency bin k. Then, the combination of the cosine and sine of the IPD spatial features are upsampled to the same frame length as the single-channel spectral representation, concatenated with the spectral (single-channel) features and fed into the separation module.

The separation module is a temporal fully-convolutional network (TCN) [Lea et al., 2016]. The TCN is built from R repetitions of a sub-block which stacks X dilated 1-D convolutional blocks. In each dilated 1-D convolutional block, the original convolution operation is replaced with a depthwise separable convolution [Chollet, 2017] to reduce the number of parameters. The separation module takes the representations from the encoder as input and estimates masks for each individual source. The masks are multiplied with the spectral representations of the mixture signal to generate separated representations. Then, the decoder reconstructs the estimated signals by inverting the separated representations back to time-domain signals. The decoder is built with a 1-D convolutional layer.

During training, a separation network generates M signals, which are compared against M targets to compute a loss function. However, the relationship between the ordering of the estimates and references is unknown. An utterance-level permutation invariant training (uPIT) criterion [Kolbæk et al., 2017] minimises the minimum utterance-level loss of all permutations and is used to train the network. Given a mixture signal $\mathbf{y}$, the separation model predicts M individual sources $\hat{\mathbf{s}}$. The uPIT loss between the estimated sources and the reference sources $\mathbf{s}$ can be written as follows:

$$\mathcal{L}_{\mathrm{PIT}}(\mathbf{s}, \hat{\mathbf{s}}) = \min_{\mathbf{P}} \sum_{m=1}^{M} \mathcal{L}(s_m, [\mathbf{P}\hat{\mathbf{s}}]_m) \tag{4.2}$$

where $\mathbf{P}$ is a permutation matrix and $\mathcal{L}$ is a signal-level loss function to be minimised. The objective function for the network is the scale-invariant signal-to-noise ratio (SI-SNR) metric (4.3), which is frequently used to assess the separation performance [Kolbæk et al., 2020]:

$$\begin{aligned} \text{SI-SNR} &= 10\log_{10}\frac{||\mathrm{s}_{target}||^2}{||\mathrm{e}_{noise}||^2} \\ \mathrm{s}_{target} &= \frac{\langle \hat{s}, s\rangle s}{||s||^2} \\ \mathrm{e}_{noise} &= \hat{s} - s_{target} \end{aligned} \tag{4.3}$$

where $\hat{s}$ and s denote an estimated source and a clean source, respectively, and $||s||^2 = \langle s, s\rangle$ denotes the signal power.

4.3 Multi-channel end-to-end separation

This section proposes a novel framework for time-domain multi-channel speech separation. The first part describes a novel approach to extract spatial features from time-domain signals. The next part describes the application of a dereverberation method as a pre-processing step for the separation task to improve feature extraction in reverberant environments. Then, an efficient network architecture is proposed to increase sequence modelling ability of the separation system. The final part describes a multi-stage separation process which applies an additional enhancement step to the separated signals.

One weakness of using IPDs is that IPDs are not well-synchronised with the learned single-channel spectral features. To extract spatial features that can be aligned with the spectral features, a neural network based encoder is proposed here. The designed end-to-end multi-channel separation system consists of a spectral encoder, a spatial encoder, a separator, and a decoder. The overall structure is illustrated in the Figure 4.2.

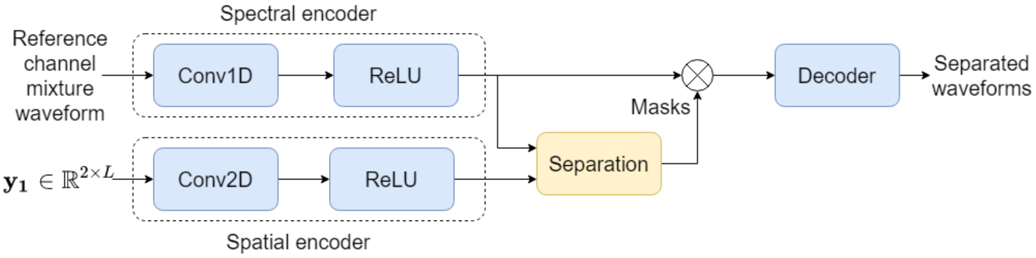


Figure 4.2: End-to-end multi-channel speech separation approach.

For the spatial information, instead of using conventional spatial features such as IPDs as additional input, a 2-dimensional (2-D) convolutional layer is used to construct the spatial encoder to extract spatial information from multi-channel recordings, keeping the whole system in an end-to-end framework. The 2-D convolutional layer has a kernel size of $(2, L)$, such that the analysis window has the same length L as the 1-D convolutional layer for the spectral encoder, thus keeping the number of frames of spatial features the same as the spectral representation. The dimension of the output channel of the 2-D convolutional layer, S, is fixed smaller than N, such that the 2-D convolutional kernel will focus more on spatial information across different channels. A non-linear activation function, i.e., ReLU, follows the 2-D convolutional layer and together they form the spatial encoder.

When the number of microphone channels, J, is larger than two, the 2-D convolutional layer can use a bigger kernel size, i.e. (J, L), to take all the channels together as input. To make a fair comparison with the IPDs which are calculated from paired signals, the proposed system continues taking a pair of signals as input and will be repeated to process all pairs when more than one pair is available. Then, the spatial representations from each pair and spectral representations from the spectral encoder are concatenated along the channel dimension and fed into the subsequent separation block.

The spatial encoder is trained jointly with the main separation network. The objective function for the multi-channel TasNet is SI-SNR, and the network is trained using uPIT.

4.3.1 Dereverberation preprocessing

When a mixture of sound sources are recorded by distant microphones in a reverberant environment, the acoustic signals can be reflected by the wall, floor, ceiling and all other surfaces in the environment. The reflected signals combine at the microphone to produce the effect of *reverberation*. The amount of reverberation will depend on the nature of the room, e.g. its size and the reflectivity of the major surfaces. It is found that the early reflections improve the sound naturalness, but the late reflections, also known as reverberation tails, deteriorate the speech intelligibility. The late reflections are a significant sources of error in automatic speech recognition even if there are no competing sources, as the reverberation itself acts as a masking noise.

There are some studies that attempt to perform joint separation and dereverberation using neural networks, in which the training targets use anechoic speech (i.e., speech recorded in conditions where there are no reflections) [Delfarah and Wang, 2019; Luo and Mesgarani, 2018a]. The single-channel time-domain network was shown to be capable of performing joint separation and dereverberation [Luo and Mesgarani, 2018a], in which the direct path component is separated out from the reverberated signal. However, Heitkaemper et al. [2020] pointed out that, for the single-channel system, the neural network based encoder is not robust to the reverberation and the quality of encoded features is low, causing a deterioration of the separated signals. In addition, for the multi-channel system, the reverberation may cause degradation to both spectral and spatial feature extraction. To address the issue, this part proposes to use an existing dereverberation method as a pre-processing step for speech separation.

The multiple input multiple output version of the weighted prediction error (WPE) method was used for dereverberation [Nakatani et al., 2010]. WPE is selected since it can conserve the spatial differences at different microphone positions, which is required for subsequent microphone array processing [Yoshioka and Nakatani, 2012]. The WPE method aims to estimate the reverberation tail of a signal and remove it from the recording to obtain a signal consisting of the direct component and its early reflections. The following part presents a brief technical description of WPE. For full details, please see Yoshioka and Nakatani [2012].

Using J microphones, a multi-channel observation in the time-frequency domain, $\mathbf{y}_{t,f}$ is observed. The observation can be split into the early component $\mathbf{x}_{t,f}^{early}$ and the reverberation tail $\mathbf{x}_{t,f}^{tail}$ as follows:

$$\mathbf{y}_{t,f} = \mathbf{x}_{t,f}^{(early)} + \mathbf{x}_{t,f}^{(tail)} \tag{4.4}$$

The WPE estimates filter weights, $\mathbf{G}_f(\tau)$, to obtain reverberation tails, and then the reverberation tails are subtracted from the observation to obtain early components. By doing so, the room impulse response for each source signal is shortened up to $\triangle$ taps.

$$\hat{\mathbf{x}}_{t,f}^{(early)} = \mathbf{y}_{t,f} - \sum_{\tau=\triangle}^{\triangle+K-1} \mathbf{G}_f^H(\tau)\mathbf{y}_f(t-\tau) \tag{4.5}$$

where K is the number of filter taps. WPE assumes that each clean speech signal follows a complex normal distribution with zero mean and time-varying variance $\lambda_{t,f}$:

$$p(\mathbf{x}_{t,f}^{(early)}; \lambda_{t,f}) = \mathcal{CN}(\mathbf{x}_{t,f}^{(early)}; 0, \lambda_{t,f}\mathbf{I}) \tag{4.6}$$

The derivation of the filter weights is based on an assumption that autocorrelation coefficients of a clean speech will be nearly zero with time lags greater than tens of milliseconds, while a reverberated signal will have larger autocorrelation coefficients with large lags. Therefore, the filter weights are achieved by minimising temporal correlation of the estimated dereverberated speech signal. There is no known closed form solution for $\mathbf{G}_f(\tau)$ and $\lambda_{t,f}$, and they need to be estimated with an iterative procedure:

$$\lambda_{t,f} = \frac{1}{(\delta + 1 + \delta)J} \sum_{\tau=t-\delta}^{t+\delta} \sum_{j}^{J} |\hat{x}_{\tau,f,j}^{(early)}|^2 \tag{4.7}$$

$$\mathbf{R}_f = \sum_t \frac{\psi_{(t-\triangle),f}\psi_{(t-\triangle),f}^H}{\lambda_{t,f}} \tag{4.8}$$

$$\mathbf{P}_f = \sum_t \frac{\psi_{(t-\triangle),f}\mathbf{y}_{t,f}^H}{\lambda_{t,f}} \tag{4.9}$$

$$\mathbf{G}_f = \mathbf{R}_f^{-1}\mathbf{P}_f \tag{4.10}$$

The new G filter estimate is used to reestimate $\hat{\mathbf{x}}_{t,f}^{(early)}$ via equation 4.5. The context of $(\delta+1+\delta)$ frames is used to estimate the time-varying variance. $\psi_{t,f}$ is a stacked representation of the observations:

$$\psi_{t,f} = [\mathbf{y}_{t,f}^H, \ldots, \mathbf{y}_{(t-K+1),f}^H]^H \tag{4.11}$$

Figure 4.3 shows the block diagram of the separation system using WPE as a dereverberation pre-processing step. Before training the separation network, the multi-channel mixtures and the individual target signals with reverberation are processed with the dereverberation method. After the dereverberation, instead of using clean anechoic sources as training targets, the separation network is trained to separate the processed mixture signal to individual dereverbed target signals. During the testing, the multi-channel mixtures are pre-processed by the WPE before being fed into the separation network. It can be noticed that this approach replaces individual anechoic signals with dereverberated signals for the training targets. This facilitates the application of joint separation and dereverberation networks to real scenarios since the matched individual anechoic signals are hard to collect in practice.

4.3.2 Improved multi-channel separation network

When modelling raw audio signals, a network should be able to model the long temporal dependencies. A CNN based sequence model takes input from a few time steps to predict

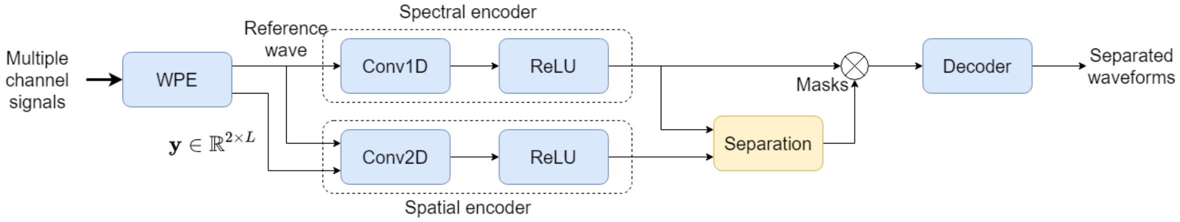


Figure 4.3: Dereverberation pre-processing for end-to-end multi-channel separation.

one output, and the region of the input for predicting the output is called receptive field for that output. For a CNN based speech separation network, a larger receptive field size can potentially lead to a better performance [Luo and Mesgarani, 2019; Tzinis et al., 2020b]. The multi-channel system developed above uses multiple stacked dilated 1-d convolutional blocks to increase the effective receptive field. However, convolutions with a large dilation factor might inject several artifacts and hurt the performance, as has been observed in a music source separation task [D'efossez et al., 2019]. An alternative approach to increase the receptive field size for convolutional architectures is subsampling, which has shown promising results for the single-channel separation task [Tzinis et al., 2020b]. Specifically, the single-channel system is constructed with a U-Net style convolutional block (U-ConvBlock), which performs iterative downsampling and upsampling operations. This part aims to apply the subsampling approach to the design of multi-channel separation system in order to increase the receptive field.

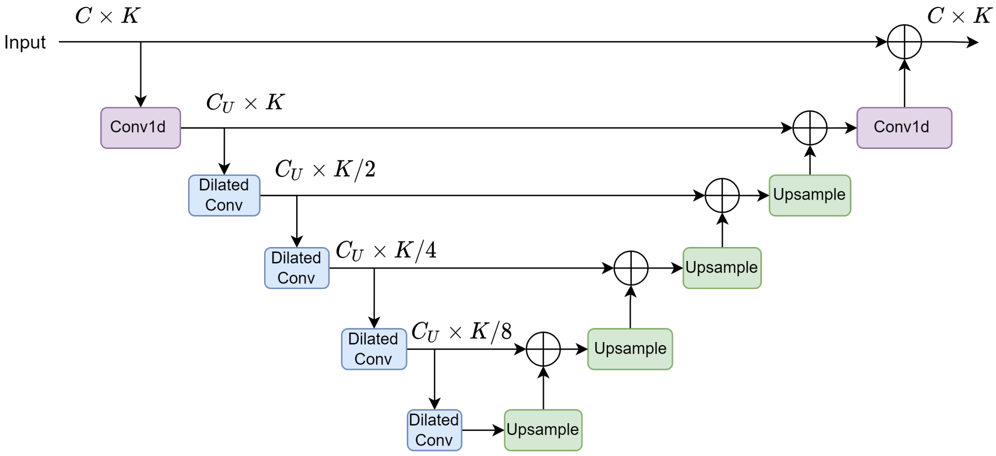


Figure 4.4: U-ConvBlock structure

The U-ConvBlock (Figure 4.4) extracts information from multiple resolutions using Q successive temporal downsampling and Q upsampling operations [Ronneberger et al., 2015]. Each downsampling operation contains a depth-wise separable convolution, and each upsampling layer contains a bilinear interpolation operation. The channel dimension of input to each U-

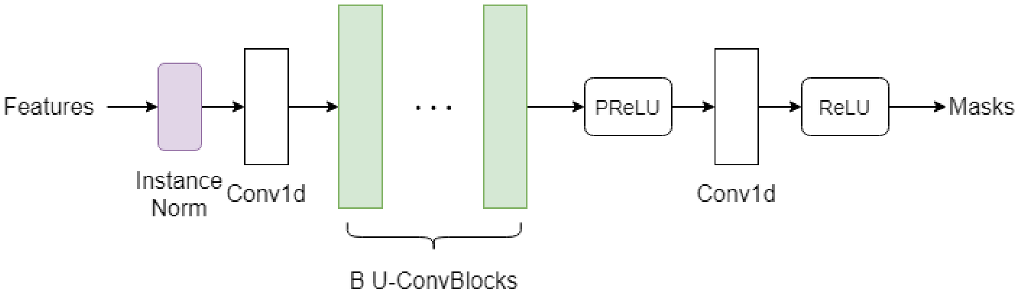


Figure 4.5: Improved separator with U-ConvBlocks

ConvBlock is expanded from C to C_U before downsampling, and is contracted to the original dimension after upsampling. The upsampled version is combined with the original, i.e. shown as the horizontal connections in Figure 4.4.

The updated separation module is shown in Figure 4.5. It consists of a layer-normalisation layer, a bottleneck layer, B stacked U-ConvBlocks and a 1-D convolutional layer with a non-linear activation function. The first layer-normalisation uses an instance normalisation layer [Ulyanov et al., 2016], in which the feature is normalised over the time dimension:

$$\begin{aligned} \mu_k &= \frac{1}{T}\sum_{t=1}^{T} f_{kt} \\ \sigma_k^2 &= \frac{1}{T}\sum_{t=1}^{T}(f_{kt}-\mu_k)^2 \\ \hat{f}_{kt} &= \frac{f_{kt}-\mu_k}{\sqrt{\sigma_k^2+\epsilon}} \odot \gamma + \beta \end{aligned} \tag{4.12}$$

where f_{kt} and $\hat{f}_{kt}$ are the input feature and normalised feature at channel k and frame t, respectively, γ and β are trainable parameters, and ϵ is a small constant for numerical stability.

4.3.3 Spatial information for decoding

Another consideration for the system design is whether or not the spatial information should be fed through to the decoding stage. In the previous system, spatial features are input to the separator only. However, features extracted from signals received by all microphones could potentially benefit the signal reconstruction in a fashion similar to conventional beamforming.

A modified system structure is proposed to take account of signals from all microphone

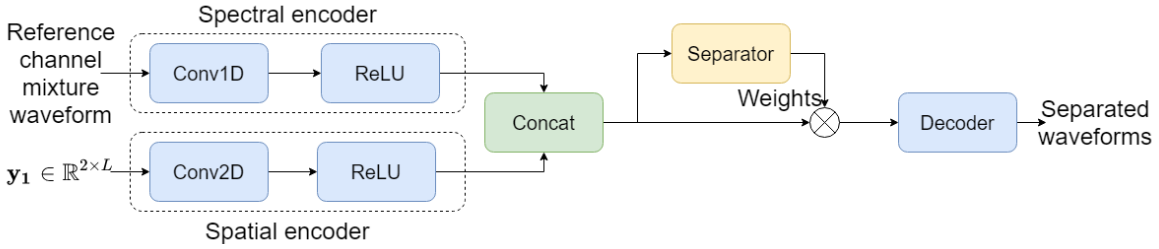


Figure 4.6: Updated multi-channel model structure. Compared to Figure 4.2, the decoder now has access to all channels.

channels during signal reconstruction. As depicted in Figure 4.6, the encoded single-channel spectral features and spatial features are concatenated together to form multi-channel representations with a dimension of $(N + S)$, which are accessed by both the separation module and the decoder. The separator will estimate linear weights for combining the multi-channel representations to generate separated representations for each source. This can also be viewed as a skip connection between the spatial encoder and decoder. The skip connection is a critical component for successful training of very deep networks because of its role in alleviating the vanishing gradient problem [He et al., 2016].

4.3.4 Multi-stage separation

Although the separation process can segregate individual signals from a mixture, it usually causes many distortions and artifacts. However, the signals can be effectively recovered from the distortions by a second enhancement stage [Isik et al., 2016]. Inspired by this, this part employs the multi-stage separation processing based on the end-to-end multi-channel system. The first stage is a standard separation stage that takes mixture signals to estimate individual sources. In the second stage, a modified multi-channel TasNet, used as an enhancement model, is trained to map the distorted signal to the clean reference signal. Figure 4.7 illustrates the structure of the enhancement system. The enhancement TasNet takes one distorted signal and the original mixture as input, and generates one enhanced output. Since the separation system in the first stage generates multiple signals, the enhancement network processes each separated signal individually.

The separation network and the enhancement network are trained separately. The first step is to train a separation network with the PIT framework. Then, the enhancement model is trained based on the separated signals generated by the pre-trained separation network, during which only the parameters of the enhancement model are updated and the parameters of the separation system are fixed.

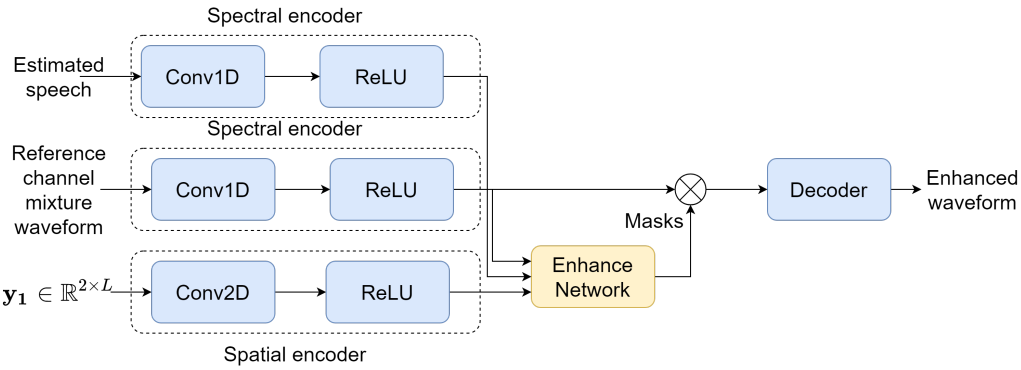


Figure 4.7: Enhancement system in the two-stage separation

4.4 Speech separation experiment setup

The proposed methods are evaluated on a sequence of tasks, starting from a speech separation task, moving forward to a joint separation and dereverberation task, and finally performing joint denoising, dereverberation and separation. For the separation task, the systems are trained with the reverberant mixture signals as input and reverberant target signals as targets. For the joint separation and dereverberation task, systems are trained to separate anechoic individual speech from a reverberant mixture. For the joint denoising, dereverberation, and separation task, the separation model is trained to separate anechoic individual speech from a noisy and reverberant mixture.

Both signal quality measurement (SI-SNR) and speech recognition performance (WER) are used as evaluation metrics. The following parts will describe the data used in experiments, configurations for speech separation network, and configurations for speech recognition evaluation. Results and discussion will then follow in Section 4.5.

4.4.1 Data

The experiments use simulated mixtures, since simulated data provides both noisy mixtures and clean reference targets, which are helpful to evaluate the speech separation performance. In addition, deep learning based separation systems require reference targets to conduct supervised training. Data recorded in real noisy environments contains only noisy mixtures and cannot be used for supervised learning. This problem of how to use noisy mixtures alone to train a separation network will be studied in Chapter 6. In this chapter, the experiments will be using two different types of data: speech mixtures with no additional noise interference, and speech mixtures with additional ambient noise. These data have been provided by two datasets commonly used for evaluation of source separation systems: spatialized WSJ0-2mix [Wang et al., 2018] and WHAMR! [Maciejewski et al., 2020].

Spatialised WSJ0-2mix [Wang et al., 2018] uses clean speech from the Wall Street Journal (WSJ) corpus and simulates multi-channel reverberant data recorded by distant microphones. The WSJ data consists of read sentences from the WSJ newspaper recorded by male and female US talkers. For each mixture utterance, a cuboid room configuration (length-width-height) is uniformly sampled from 5-5-3 m to 10-10-4 m, as shown in Figure 4.8. In the centre of each room is located a circular array with 6 elements arranged uniformly with an angle of separation of 60 degrees. The radius of the array is uniformly drawn from a range between 7.5 cm and 12.5 cm. Two different speakers are randomly positioned in a rectangular area that has 3 m length and 3 m width in the centre of the room. Then the classic image method [Allen and Berkley, 1979] is used to simulate multi-channel room impulse responses with T_{60} (i.e., the degree of reverberation) uniformly distributed between 200 ms and 600 ms. The multi-channel signals are obtained by convolving the clean speech with the room impulse responses. Two speakers' speech is mixed using a pre-defined signal-to-noise ratio value, which ranges from absolute -2.5 dB to 2.5 dB. Note, to match the exact SNR values of the anechoic mixture, the reverberant source images are rescaled.

It should be noted that, this scenario is not realistic and many factors in real environments are not considered. For example, the location of each speaker in a mixture is fixed; directivity patterns for human speech are not considered; reflections of speech signals only come from walls, floors, and ceilings; and there is no background noise or diffuse noise sources. However, the dataset has become widely used and provides a useful standard with which to compare approaches, under the assumption that their relative performance on this simpler data is an indicator of potential performance on real data.

WHAMR! dataset [Maciejewski et al., 2020] simulates stereo (2-channels) recordings and

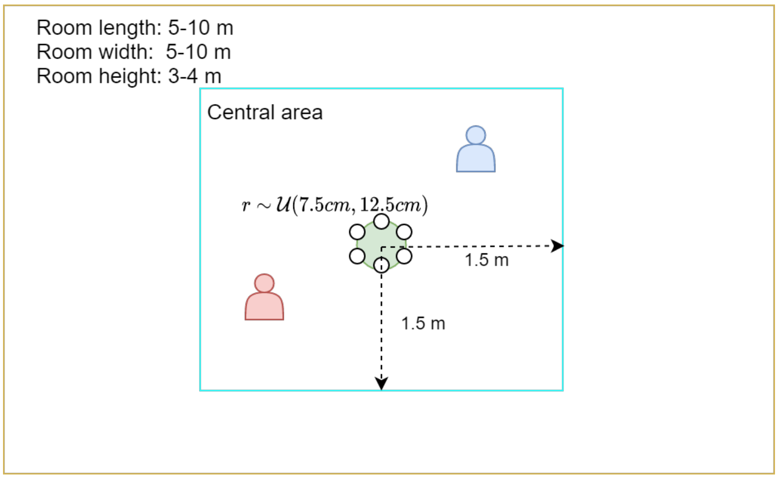


Figure 4.8: Spatialised WSJ0-2mix simulation illustration

Table 4.1: WHAMR! Room impulse response parameter configurations

Room	L (m)	$\mathcal{U}(5, 10)$
	W (m)	$\mathcal{U}(5, 10)$
	H (m)	$\mathcal{U}(3, 4)$
$\mathbf{T_{60}}$	low (s)	$\mathcal{U}(0.1, 0.3)$
	med. (s)	$\mathcal{U}(0.2, 0.6)$
	high (s)	$\mathcal{U}(0.4, 1.0)$
Microphone Center	L (m)	$\frac{L_{Room}}{2} + \mathcal{U}(-0.2, 0.2)$
	W (m)	$\frac{w_{Room}}{2} + \mathcal{U}(-0.2, 0.2)$
	H (m)	$\mathcal{U}(0.9, 1.8)$
Source Locations	H (m)	$\mathcal{U}(0.9, 1.8)$
	dist. (m)	$\mathcal{U}(0.66, 2)$
	angle	$\mathcal{U}(0, 2\pi)$

is also based on the WSJ clean speech corpus. This simulation can better approximate the real world scenario, since it not only considers simultaneous speech and reverberation, but also background noise. The noise is collected by using an Apogee Sennheiser binaural microphone in various urban environments consisting of restaurants, cafes, bars, and parks [Wichern et al., 2019]. The microphone with an inter-microphone distance between 15-17 cm was mounted on a tripod positioned on a table with heights ranging between 1.0-1.5 m. The same inter-channel distance is used by the *pyroomacoustics* toolkit [Scheibler et al., 2018] to simulate room impulse responses. For each utterance, both anechoic and reverberant versions are simulated. A cuboid room configuration (length-width-height) is uniformly sampled from 5-5-3 m to 10-10-4 m. The reverberation times T_{60} range from 100 ms to 1 second, and are further classified as low, medium, and high reverberation, in which the T_{60} is uniformly sampled from 100 ms to 300 ms, 200 ms to 600 ms and 400 ms to 1 second, respectively. The room configurations are summarised in Table 4.1. Two speakers' speech is mixed using a pre-defined signal-to-noise ratio, which ranges from absolute -2.5 dB to 2.5 dB. The reverberant speech is mixed with the recorded noise such that the louder speaker is at a randomly selected SNR between -6 dB to 3 dB relative to the noise.

For both spatialised WSJ0-2mix and WHAMR!, The size of the training, validation, and test sets are 20k, 5k, and 3k utterances, respectively. The sampling rate of all the data is 8 kHz. The speakers in the validation set are the same as in the training set, but the speakers of the utterances used for testing do not appear during training, i.e., the system is being evaluated in a speaker-independent setting. The training and test sets contain 101 and 18 speakers, respectively. For the WHAMR! dataset, each of the training, validation, and test set contains noise recordings from all four environments. Data with no reverberation, but with identical microphone and source geometry, has also been simulated to show the separation performance in the anechoic environment and for use in the dereverberation experiments.

4.4.2 Speech recognition evaluation

The speech recognition performance is evaluated with the long established WSJ corpus. The acoustic models have a topology with a 12-layered Factorised TDNN (TDNN-F) [Povey et al., 2018], which can produce state-of-the-art performance. Each layer has 1024 units. The input to the acoustic model is 40-dimensional MFCCs and a 100-dimensional i-Vector. The acoustic models are trained with Lattice-Free Maximum Mutual Information (LF-MMI) [Povey et al., 2016] and are implemented with the Kaldi speech recognition toolkit [Povey et al., 2011]. A 3-gram language model is used during recognition. The audio data is downsampled to 8 KHz to match the sampling rate of data used for the separation experiments.

Two acoustic models have been trained with different training strategies. One model (AM1) was trained on roughly 80 hrs of clean WSJ0/WSJ1 SI-284 data, i.e. the standard WSJ training set. The second one (AM2) used a multi-condition training strategy and was trained on the same clean data plus a spatialized version of it (in total roughly 160 hrs of training data). Randomly selected room impulse responses used to simulate the multi-channel mixtures were employed to reverberate the train set of AM2. A simple 3-gram language model is used for decoding. With this set-up the ASR results obtained with AM1 on the standard clean WSJ Dev93 and Eval92 are 7.2% and 5.2% WER, respectively.

4.4.3 Separation network configuration

The single-channel Conv-TasNet [Luo and Mesgarani, 2019], SuDoRM-RF [Tzinis et al., 2020b], and multi-channel system IPD-TasNet [Gu et al., 2019] are used as baseline systems for comparison. For the single-channel system, the hyper-parameters of Conv-TasNet are set as those that produced best performance in the original paper [Luo and Mesgarani, 2019], namely, $N = 256$, $R = 3$, $X = 8$, $L = 16$, and the batch size $M = 3$. For the single-channel SuDoRM-RF, the hyper-parameters are set to match those of SuDoRM-RF 1.0x in Tzinis et al. [2020b], namely, $L = 17$, $B = 16$, $Q = 4$, $C = 256$, and $C_U = 512$.

The IPD-TasNet with fixed kernels in Gu et al. [2019] has been reproduced as the multi-channel baseline system. For the IPD spatial feature calculation, the window length of the STFT is set to 64 samples. During implementation, the 1-D convolution kernel that encodes multi-channel information is initialized by the STFT parameters and then fixed during the training. The spatial feature input to the separation block is $\cos(\text{IPD}) + \sin(\text{IPD})$, with a dimension of 33.

For the proposed multi-channel separation system, the separation module in the TCN based system shares the same configuration to the single-channel Conv-TasNet. In the spatial feature extraction experiment, the number of filters, S, in the 2-D convolutional layer is set to 36, similar to the dimension of the IPD feature, to make a fair comparison between the learned features and the conventional features. The window length of the 2-D convolutional layer is the same as the 1-D convolutional layer used in spectral encoder. For the multi-stage framework, the second-stage enhancement system uses the TCN based structure and the parameters are set as $N = 256$, $R = 3$, $X = 8$, and $L = 16$.

All the separation models are trained using the Adam optimizer [Kingma and Ba, 2014] with a learning rate of $1e-3$, which is halved every time the loss of validation set is not reduced in 3 consecutive epochs. All models are trained with 100 epochs. The networks are trained on four seconds long speech segments which are obtained by splitting each utterance. These training configurations are set as those that are used to train the Conv-TasNet [Luo and Mesgarani, 2019].

4.5 Results and analysis

This section presents the experiments used to evaluate the proposed methods and analyses the results. The first part evaluates the neural network based spatial feature extraction for multi-channel separation. The second part investigates the effect of reverberation on end-to-end separation and shows the results of the proposed approach that applies a dereverberation as pre-processing. Then, the improved separation network architecture is tested. The final part shows the performance for the multi-stage separation procedure.

4.5.1 Improved spatial features

The spatial feature extraction methods are evaluated with two- and six-channel separation, based on different microphone pair selections. The microphone pairs are made by selecting both nearest and furthest microphones, as shown in Figure 4.9. The experiments with 2 channels use the pair (1, 4), on opposite sides of the circular array, such that the distance between the selected microphones equals the array's diameter. Pairs of signals, (1, 4), (2, 5), (3, 6), (1, 2), (3, 4) and (5, 6), are selected for all multi-channel experiments with 6 channels.

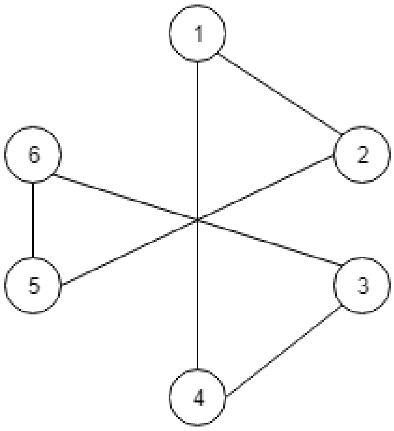


Figure 4.9: Microphone pairs selection from the circular array

Results are shown in Table 4.2. We first evaluate the baseline single- and multi-channel systems in both anechoic and reverberant conditions. In the anechoic condition, the single-channel TasNet achieves 14.6 dB SI-SNR improvement, matching the reported results in Luo and Mesgarani [2019]. The performance of the single-channel TasNet drops noticeably (nearly 8 dB) in the reverberant case, indicating that reverberation has a strong impact on the source separation performance of standard TasNet. The IPD based multi-channel system achieves 19.7 dB and 10.9 dB in anechoic and reverberant conditions, respectively, which represents significant SI-SNR improvements in both conditions compared with the plain single-channel TasNet. This indicates that the spatial information benefits the end-to-end separation, although the spatial features and spectral features are extracted from different signal domains.

Table 4.2: SI-SNR improvements of reference and proposed systems.

System	#chs	SI-SNRi (dB)	
		Anechoic	Reverb
TasNet [Luo and Mesgarani, 2019]	1	14.6	6.7
IPD-TasNet [Gu et al., 2019]	6	19.7	10.9
Proposed	2	28.2	10.9
Proposed	6	**30.0**	**12.6**

Then, the speech separation performance of the proposed spatial feature extraction has been evaluated. The proposed speech separation method that uses 2 channels has either outperformed or matched the performance of the 6-channel IPD system in the anechoic and reverberant conditions, respectively. Using 6 channels led to almost perfect anechoic separation and yielded more than 13% relative SI-SNR improvement over the reference IPD system and the proposed system with 2 channels in reverberation, suggesting that the fully convolutional structure can benefit from more microphone channels. One possible reason for the superiority of the proposed system over the one based on the IPD could be that the proposed spatial features extracted using 2-D convolutional layers are aligned with the learned spectral features from the encoder, whereas for IPD extraction the STFT window length is larger than the window length L of the convolution kernel, therefore the IPD features need upsampling which may cause misalignment with the spectral features.

Figure 4.10 illustrates the learned filters of the 2-D convolutional layer. The learned filters calculate the difference between two signals, demonstrating that the spatial encoder learns to find correlation between two microphone channels instead of encoding single-channel information from each channel. The learned basis functions have different valid window lengths, which are all much shorter than the normal window length of the STFT, indicating the network itself can find better filters to encode the spatial features.

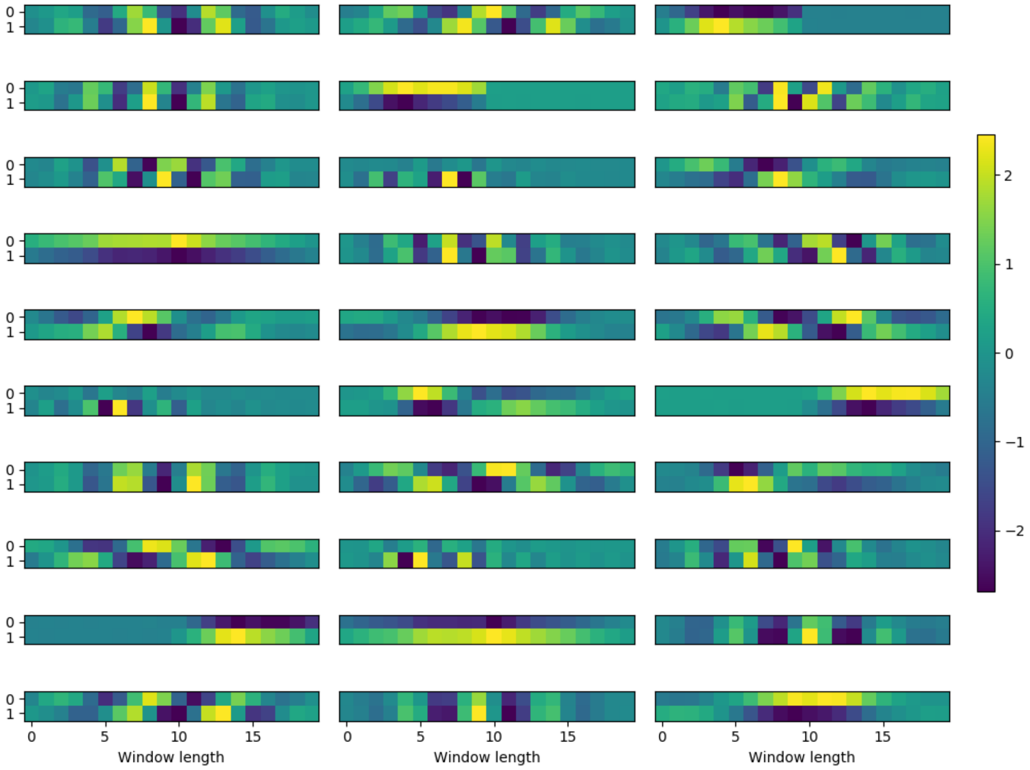


Figure 4.10: Basis functions of 2-D spatial encoder of proposed method.

We now wish to test the speech recognition performance by using the speech separation method as front-end processing. The evaluation sets are first passed through the source separation, then both the separated outputs are decoded by the ASR system. The WERs are computed according to the reference transcriptions.

The results are depicted in Table 4.3. The unprocessed mixture in both anechoic and reverberant conditions yield very high WERs, around 80%. This indicates the difficulty for an ASR system when recognising overlapping speech. In the anechoic condition, the WER of the single-channel TasNet is around 19%, which is significantly better than the results based on GMM-HMM [Isik et al., 2016] or end-to-end ASR models [Chang et al., 2019; Seki et al., 2018]. Remarkably, all the multi-channel separation models are able to achieve a speech recognition performance close to the Oracle result in the anechoic condition. The acoustic with multi-condition training (AM2) achieves similar accuracy compared with the model trained on clean

Table 4.3: ASR accuracies in WER(%) of reference and proposed systems for anechoic and reverb conditions.

System	#nchs	Anechoic		Reverb	
		AM1	AM2	AM1	AM2
Mixture	1	79.6	80.8	88.6	82.4
TasNet [Luo and Mesgarani, 2019]	1	19.3	18.5	74.5	47.1
IPD-TasNet [Gu et al., 2019]	6	10.6	10.3	70.0	40.3
Proposed	2	10.7	10.6	62.2	25.3
Proposed	6	9.2	9.2	57.2	19.8
Oracle	1	9.1	9.1	46.2	11.3

speech signals (AM1), since the signals are well separated and there is no reverberation.

In the reverberant case, the proposed multi-channel separation system provides both acoustic models with consistent performance gains over the IPD system. Notably, for AM2, the proposed system with 6 channels yields more than 50% relative WER improvement over the IPD system with 6 channels. An interesting observation is that although the SI-SNR improvement in Table 4.2 for the reference IPD system and the proposed 2-channel method in reverberation are similar (both 10.9 dB), the corresponding WER results in Table 4.3 are significantly different, i.e., 40.3% for the IPD system and 25.3% for the proposed system. This observation requires further analysis in the future, and it may show that the SI-SNR metric may not be well-motivated when deploying speech separation enhancement for multi-talker ASR.

AM2 performs much better than AM1, since AM2 is trained with multi-condition data, thus is more robust to the distortions introduced by the reverberation. However, the ASR gain from the separation pre-processing in the reverberant condition is still less than that in the anechoic condition. And the ASR performance of separation in the reverberant condition is still much worse than the performance of the oracle reverberant speech signals (11.3% of WER), suggesting that the system has difficulties in extracting spectral features and spatial features from signals with reverberation. Next, the effect of dereverberation and speech separation on the ASR accuracy is investigated.

4.5.2 Dereverberation preprocessing

Three different approaches for joint separation and dereverberation are investigated. Firstly, the reverberant unprocessed speech mixtures and the clean (anechoic) targets were used to train the separation systems, thus constraining the networks to learn both to separate the speakers and perform dereverberation. Both the single-channel system and the multi-channel system have been trained in an end-to-end fashion. These systems are provided to allow comparison with the system using WPE as a preprocessing step, showing the problem of extracting spatial features directly from reverberant signals. Secondly, the mixtures are processed with the WPE. The dereverberated mixtures and anechoic targets were employed for training. This was to investigate if reverberation in the mixture impacts the spectral and spatial feature extraction in the end-to-end separation approach. Lastly, we consider a situation where anechoic targets are not available. In this case, both the mixtures and the reverberant targets were enhanced using WPE, and the dereverberated signals are used for training the separation system.

Results are summarised in Table 4.4. 'None' denotes that the reverberant signal is unprocessed. 'Clean' denotes that the anechoic signal is used. 'WPE' denotes that the reverberant signal is processed with the WPE processing. The results for the single-channel TasNet show that there is a consistent WER improvement with both acoustic models for the None-Clean case (second row) compared with the None-None case (first row). This demonstrates that the system is able to jointly separate and dereverberate the signals by using anechoic speech as training targets for the separation model. But the ASR accuracy is lower than in the WPE-Clean case (third line), demonstrating that the reverberation has a negative impact on the separation performance and the separation model can extract better spectral representations from a dereverbed mixture.

Concerning the proposed 6-channel separation system, similar trends can be observed, but the absolute improvements are significantly larger. In this case, for the WPE-Clean condition there is a consistent WER improvement with both acoustic models compared with the None-Clean case, indicating that a more powerful spectral and spatial signal representation can benefit from cleaner mixtures during training, i.e., the network is able to remove the mild reverberation remained in the WPE processed data. The acoustic model trained with reverberant data (AM2) outperforms the model trained only with clean speech (AM1), however,

Table 4.4: Results of different source separation and dereveberation strategies. Targets' enhancement is during training only.

System	Enhancement		WER(%)		SI-SNRi (dB)
	Mixture	Targets	AM1	AM2	
TasNet [Luo and Mesgarani, 2019] (1-ch)	None	None	74.5	47.1	6.7
	None	Clean	53.3	44.1	8.7
	WPE	Clean	38.4	31.8	8.8
	WPE	WPE	39.0	30.3	9.6
Proposed (6-ch)	None	None	57.2	19.8	12.6
	None	Clean	25.2	20.0	15.2
	WPE	Clean	16.3	14.0	16.0
	WPE	WPE	20.0	14.4	15.5

applying dereverberation has reduced the performance gap between the models.

Finally, it can be observed that applying WPE to reverberant targets during the training of the separation model yields almost the same WER as the WPE-Clean condition for AM2. For the single-channel separation system, the performance in the WPE-WPE case is 30.3% for AM2, similar to the result 31.8% achieved in the WPE-Clean case. And for the multi-channel separation system, the performance in the WPE-WPE case is 14.4% for AM2, which is almost same to the result as the 14.0% achieved in the WPE-Clean case. Therefore, competitive WERs can be obtained without accessing the anechoic individual speech, which makes it easy to apply the separation method in real applications.

These results indicate that combining spectral and spatial signal representations in an end-to-end fashion helps improve speech separation and ASR accuracy. Also, dereverberation pre-processing can yield significant performance improvement.

4.5.3 Improved multi-channel separation network

The evaluation of the improved architecture is performed on WHAMR! dataset, doing a joint denoising, dereverberation, and separation task. The network is trained with noisy and reverberant mixture as input and clean speech as targets.

Table 4.5 reports the separation performance for the improved multi-channel separation

Table 4.5: The effect of different configurations for end-to-end multi-channel separation. L, S, and B denotes the encoder window length, the spatial feature dimension, and the number of repetitive U-ConvBlocks, respectively.

Model	**L**	**S**	**B**	**SI-SNRi**
Multi-TCN	20	36	-	12.1
Multi-UConv	21	36	16	12.7
Multi-UConv	17	36	16	12.8
Multi-UConv	17	64	16	12.8
Multi-UConv	17	128	16	12.8
Multi-UConv	17	256	16	12.8
Multi-UConv	17	256	20	12.9
Multi-UConv (spatial decoding)	17	256	20	13.1

network with different configurations, including the building block in the separation module, the encoder window length L, spatial feature dimension S, number of repetitive blocks in the separation module B, and skip connections of the spatial encoder. The first observation is that replacing the TCN blocks with the stacked U-ConvBlocks yields 0.6 dB SI-SNR improvement. This is believed to be because the U-ConvBlock architecture employs the temporal sampling operations to increase the receptive field size of the network, while maintaining the temporal resolution intact. The second observation is that reducing the window length from 21 samples to 17 samples yields a slightly better performance, i.e., 0.1 SI-SNR gain. This is possibly due to the fact that using a smaller window length for the encoder can encode the raw audio signal into representations with a higher resolution. Increasing the dimension of the spatial encoder has not shown benefits to the separation performance. Finally, increasing the number of repetitive blocks B from 16 to 20 leads to a deeper network and increases the model capacity, showing a further improvement.

The last row in Table 4.5 shows the performance achieved by the system that uses the spatial features for signal reconstruction. The proposed method provides another 0.2 dB gain of SI-SNR compared to the model that only decodes single-channel representations from the reference channel. The result indicates that, for the end-to-end multi-channel separation, the decoder should use features from all microphone channels.

Since the speaker location in each mixture is uniformly sampled, the angular separation

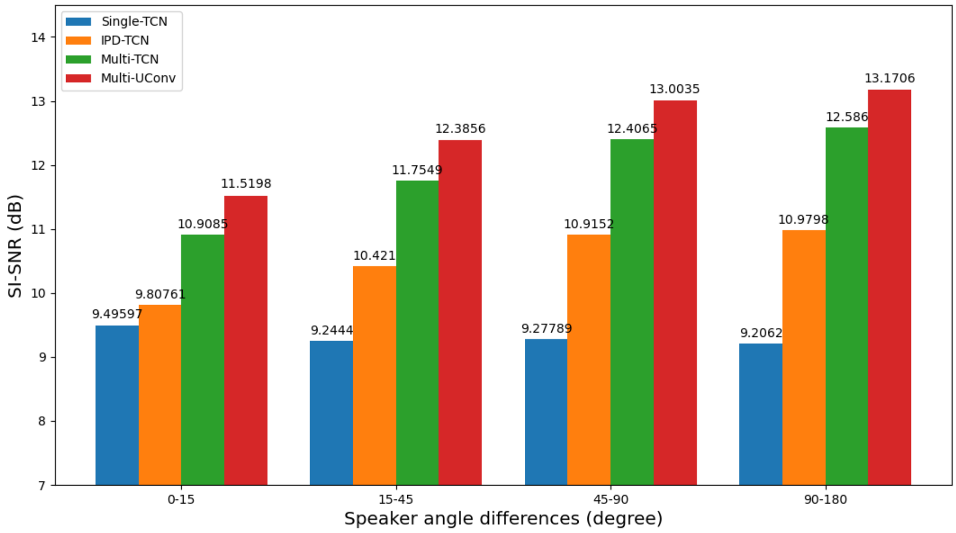


Figure 4.11: SI-SNRi performances of referenced and proposed methods on different speaker angle conditions

between two speakers in a mixture varies during the simulation. To understand the effect of the angle of separation between two speakers on the performance of multi-channel systems, the test set is split into four subsets specified by the speaker angular distance in each mixture. The samples with angular distance between two simultaneous speakers of $0-15°$, $15-45°$, $45-90°$, and $90-180°$ respectively account for 9%, 17%, 24%, and 50% of the data. The results are illustrated in Figure 4.11. It can be observed that the single-channel model achieves similar results in all conditions.

The IPD-TasNet improves the performance if the speakers are spatially separable, but provides limited benefits when the angle difference is small. The proposed Multi-TCN system outperforms the reference IPD-TasNet in all conditions. Specifically, the Multi-TCN not only provides greater gains than the reference system when the speakers are spatially separable, but also benefits the separation when the spatial difference is relatively small. This clearly shows that using a learnable kernel to extract spatial features is more suitable than conventional STFT based features for an end-to-end separation system. Furthermore, replacing the TCN block with the U-ConvBlocks in the multi-channel separation model improves the performance in all conditions. This result suggests that the sequence modelling capacity is a key factor of the design of the end-to-end multi-channel separation model.

4.5.4 Multi-stage separation

The second column in Table 4.6 shows the speech separation results with the enhancement post-processing. When using the Multi-TCN as the first-stage separation model, the enhancement post-processing provides an SI-SNR gain of 1.2 dB.

To evaluate the effect of the multi-stage separation on the speech recognition performance, an acoustic model was trained on the clean WSJ0/WSJ1 SI-284 data plus the WHAMR! single noisy reverberant speech. The overall training set contains roughly 200 hours data. The acoustic model topology is the same as AM1 and AM2 that were described in Section 4.4.2. A 3-gram language model is used during recognition.

The last column in Table 4.6 shows the speech recognition performance. The multi-stage separation process not only improves the quality of separated signals, but also the speech recognition accuracy. The WER is reduced from 39.3% to 35.1%. It is also observed that when the separated signals have better quality, the enhancement model can take advantages of this and achieve better performance as well. Using the Multi-UConv as the first-stage separation model yields better second-stage results than using the TCN based model for the first-stage separation. For the speech recognition performance, a relative 10% WER reduction is observed.

Table 4.6: Results with second-stage enhancement

Method	**SI-SNRi (dB)**	**WER (%)**
Multi-TCN	12.1	39.3
+Enhancement	13.3	35.1
Multi-UConv	13.1	34.9
+Enhancement	13.8	31.6

4.6 Summary

In this chapter, a new framework for end-to-end multi-channel separation has been developed. A 2-D convolutional layer is used to construct a spatial encoder to effectively extract spatial information from multi-channel time-domain signals. The spatial encoder learns to find correlation between two microphone channels and extracts spatial features with various valid window lengths. This end-to-end approach efficiently addresses the misalignment and mismatch problems which degrades the performance when incorporating time-frequency domain spatial features into the end-to-end system. The learned spatial features can be better aligned and combined with the spectral features, leading to a better multi-channel separation performance.

A further investigation shows that, in realistic environments, reverberation degrades the spectral and spatial feature extraction of the end-to-end separation model. Applying dereverberation methods to the mixture signals as a preprocessing step, allows better features to be extracted by the separation system, leading to a better speech separation performance in reverberant environments. The improved quality of the separated signals can also benefit a subsequent ASR system.

To increase the effective receptive field of a multi-channel separation system, dilated convolution and subsampling for convolutional networks have been compared. The results show that successive downsampling and upsampling operations are more effective at increasing the receptive field while introducing fewer distortions to the temporal resolution in comparison to the dilated convolution with a large dilation factor. The encoder-decoder architecture for the end-to-end multi-channel system is further investigated. It is argued that the decoder should use both spectral and spatial features for signal reconstruction, rather than using spectral features from the reference channel alone. A detailed analysis in different conditions of angular separation shows that the proposed spatial extraction method and the modified architecture benefit the separation in both small and large angle difference conditions.

Finally, a multi-stage separation procedure is investigated. Using an enhancement system as a post-processing step effectively reduces the distortion caused by the separation process, resulting in a higher signal quality of separated speech compared to the non-enhanced signal. The enhanced signals also achieve better speech recognition performance compared with separated signals from the single-stage separation.

To this end, a multi-channel time-domain separation system has been designed. However, there are some limitations to this separation system. One is that the speaker identity of each separated signal is unknown. In this chapter, the identity is estimated by accessing references, which is not practical. Another limitation is that the design of a separation system requires prior knowledge of number of sources in a mixture. In the separation design, the number of outputs of a separation system usually equals to the number of sources. Therefore, such a separation system cannot perform well in scenarios where number of speakers varies at different time. The third limitation is that training a separation network requires both noisy mixtures and corresponding reference targets. Chapter 5 attempts to address the first two limitations. The third limitation is considered in Chapter 6.

Chapter 5

Speaker-conditioning for time-domain speech separation

5.1 Introduction

In the previous chapter, it was shown that an end-to-end multi-channel separation system is able to segregate individual sources with a good quality. The separation model can successfully serve as a front-end processing stage for an automatic speech recognition system and can improve the recognition performance during periods of overlapping speech. However, there are three major issues that limit the applicability of the proposed separation approach. First, the design of the speech separation network requires prior knowledge of the total number of speakers in a mixture. Second, the speaker identity of the separated signals during inference is still unknown. The subsequent ASR system relies on the speaker identity to select and recognise the speech from a target speaker. The identity information can be obtained by using a speaker recognition system but this could introduce additional identification errors. Third, clean individual sources are hard to collect in a real scenario, especially when only distant microphones are used to record speech signals. This issue makes it difficult to conduct supervised learning for training a separation network.

In recent years, an alternative approach to segregate individual speech from a mixture has emerged, which is target speaker extraction [Delcroix et al., 2020; Ge et al., 2020; Žmolíková et al., 2019]. In this case, the separation model is provided with information about the identity of the target speaker to extract from the mixture. Since only one speech signal is to be estimated and the identity of that speaker has been provided, there is no permutation ambiguity. In addition, the design of an extraction system does not need prior knowledge of the number of sources in a mixture. Therefore, a speaker extraction system can be applied to situations where the number of simultaneous speakers varies.

The speaker information can be provided in many different forms, including visual information [Ephrat et al., 2018], speaker locations [Chen et al., 2018b], or voice characteristics [Delcroix et al., 2018]. The speaker information should distinguish between the target and interference speakers so that it helps an extraction system identify the target speaker in a mixture and recover the speech of this speaker. The question of what type of speaker information to employ may depend on the specifics of the situation. For example, voice characteristics between two different genders are more distinct than characteristics between two same genders. And speaker location information will only work well if the speakers are located at different positions.

Most extraction systems only exploit identity information from a target speaker and aim to recover speech from this single speaker. In other words, this framework ignores the interfering sources in the mixture. Using the additional speaker information has been shown to improve the quality of the target signals compared to separation approaches without exploiting speaker identity representations [Delcroix et al., 2020; Žmolíková et al., 2019]. However, if the prior knowledge of the distractor is also available, it remains unknown whether this knowledge can benefit the extraction of the target speaker. In addition, we may be interested in multiple speakers in a mixture. To extract multiple speakers in a mixture, the current extraction system that targets a single speaker needs to be applied several times, which causes inconvenience. Therefore, it is of practical importance to develop an extraction system that is capable of exploiting identity information from multiple speakers to extract speech from corresponding speakers simultaneously.

Strategies concerning how to incorporate speaker information to the extraction system play a critical rule for the design of an extraction system. Common strategies include multiplication [Delcroix et al., 2020] and concatenation [Ge et al., 2020] operations. However, these mechanisms may not effectively combine features from multiple modalities when dealing with the multiple tasks of speaker identification and speech reconstruction in the speech extraction system. When the extraction system is designed for multi-channel recordings, this problem becomes harder since the features to be combined come from three modalities, including spatial features, spectral features, and speaker identity features. If they are not combined properly, features from each modality may not be fully exploited [Delcroix et al., 2020; Zorila et al., 2021]. The question of how to effectively exploit the speaker information for a speaker extraction system needs further study.

A speaker extraction system relies on the quality of the speaker identity representations. When voice characteristics are used as speaker information, the identity information is usually represented as an embedding vector that can be extracted from enrollment sentences uttered by the speaker of interest. Previous studies only use one enrollment sentence to generate the speaker identity representation. However, selecting a single utterance for representation generation does not guarantee that the speaker embedding is of good quality and may lead to unstable extraction performance [Li et al., 2019]. Using multiple utterances is expected to improve the speaker representation and lead to better extraction performance.

This chapter aims to develop a speaker extraction system by extending the blind multi-channel separation system presented in the previous chapter. To improve the robustness of speaker representations, multiple enrollment utterances are exploited. The extracted representations from each utterance are averaged to form a global representation for each speaker to stabilise the extraction system. To include speaker information into the extraction system effectively, a novel conditioning mechanism has been designed to coordinate static speaker embedding with sequential signals. The proposed conditioning mechanism introduces a novel speaker stack branch to receive speaker identity features. With the proposed conditioning mechanism, a further study investigates, for a multi-speaker extraction system, if the knowledge of the distractor speaker can benefit the extraction of the target speaker.

This chapter starts with a description of the essential background knowledge for building a speaker extraction system in Section 5.2, and then considers the design of a time-domain multi-channel extraction system in Section 5.3. Implementation details and the experimental setup are described in Section 5.4. Section 5.5 presents the experiments and results. Section 5.6 summarises the work in this chapter.

5.2 Background

In general, a speech extraction system consists of two networks, one to generate speaker embeddings, and another one to perform speech extraction. As illustrated in Figure 5.1, the extraction network is conditioned on the target speaker identity information represented as a speaker embedding vector which enables it to extract the target speech given a mixture. The system is trained to minimise a signal quality loss, e.g., SI-SNR, between the estimated signal and the target signal. For most speaker extraction systems, only one target speech signal is generated, and there is no label ambiguity. The following parts will describe in detail the generation of speaker embedding, the network architecture of the extraction model, and the exploitation of multi-channel signals.

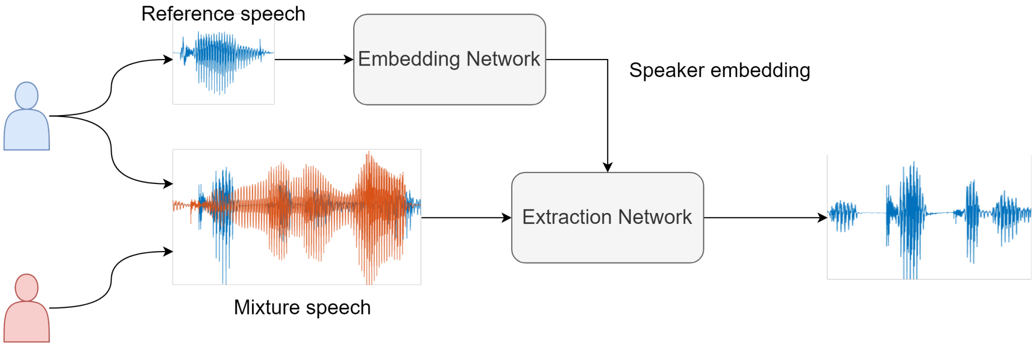


Figure 5.1: The general structure of a speaker extraction system. There may be one of more distractor speakers.

5.2.1 Speaker embedding

The speaker embedding can be either obtained from a pre-trained speaker recognition model [Wang et al., 2019b] or learned internally with an auxiliary network included in the speech extraction system [Delcroix et al., 2018]. For the pre-trained approach, a speaker recognition or verification model is usually trained to optimise a speaker classification objective, and one hidden layer inside the model is selected to generate the speaker embedding. For example, a speaker extraction system developed in Wang et al. [2019b] uses a pre-trained speaker verification model to generate a speaker embedding vector with a fixed dimension of 256. The speaker verification model is constructed by a 3-layer LSTM network and is trained with a generalised end-to-end loss [Wan et al., 2018]. The verification model takes as input a 1600 ms segment to estimate one embedding vector. To obtain an embedding vector for one utterance, the utterance is split into short segments with 50% overlap and the obtained embedding vectors for each segment are L2-normalised and averaged.

Externally training a speaker recognition model can enable the use of large corpora designed for the speaker verification task, which usually has larger speaker variations compared with the data used for training a speech extraction system. This can improve the speaker discrimination of the embedding. However, since the embedding is trained to optimise an objective function related to the speaker classification task, the obtained embedding may not be optimal for the speaker extraction task [Ji et al., 2020].

For the jointly learned speaker representation, an auxiliary network is trained with the main extraction network to minimise signal construction loss between the estimated signal and the reference speech. Although the auxiliary network is not trained with a speaker-classification related objective, it has been shown that the output of the auxiliary network can capture speaker identity information [Žmolíková et al., 2019]. In this way, since the speaker information extracted from the enrollment sentence is learnt jointly with the extraction network, the speaker representation is more suitable for the extraction task. The auxiliary network can be formed with multiple trainable layers, e.g., CNNs or LSTMs. On top of the auxiliary network, there is a sequence summary operation to map sequential features to a single speaker embedding vector [Veselý et al., 2016].

There are studies that exploit multi-task learning of speaker identification and speaker extraction [Delcroix et al., 2020]. In this case, the learned speaker embedding from the the auxiliary network is also used for a speaker identification task, which is optimised jointly with the speaker extraction task. It was shown that the speaker identification task can train the auxiliary network to obtain more discriminative speaker embedding vectors, which would help the extraction network to separate the target speaker from a mixture. Figure 5.2 shows the general framework for the jointly learned embedding and extraction approach with the multi-task training.

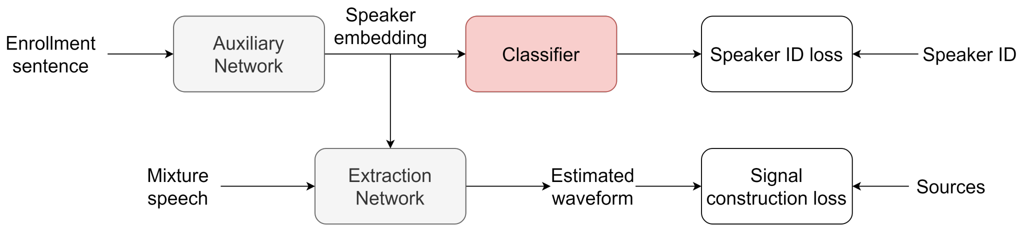


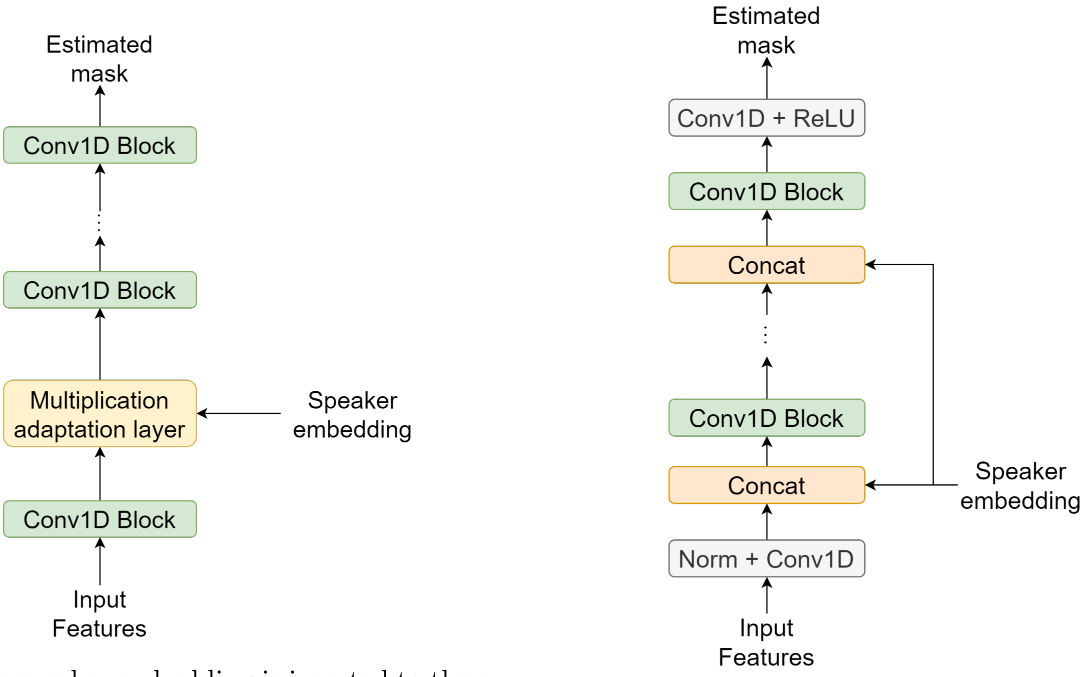
Figure 5.2: The jointly learned approach with multi-task framework of speaker identification and speaker extraction

5.2.2 Speech extraction network

Most of the earliest-published deep-learning based extraction systems operate on the time-frequency (T-F) domain [Delcroix et al., 2018; Wang et al., 2019b; Žmolíková et al., 2017]. They take T-F domain mixture signals and the target speaker representation as input to estimate T-F domain related signals for the target speaker. The T-F domain representations are obtained by applying the short-time Fourier transform (STFT) to time-domain signals. The training target of an extraction system can be either a T-F mask for the target speaker or the clean target signal in the T-F domain. Then, the inverse Fourier transform is used to transform the estimated T-F signal back to time-domain.

Inspired by the success of the time-domain separation system, recently developed speaker extraction systems have replaced the STFT and its inverse operations with trainable encoder and decoder stages to directly operate on time-domain signals [Xu et al., 2020a, 2019b]. A time-domain speaker extraction network usually consists of three components: an encoder, an extractor, and a decoder. The encoder uses a convolutional neural network to transform the time-domain mixture signals into N-dimensional representations. The extractor can be built by stacking multiple convolutional blocks [Xu et al., 2019b]. It takes the encoded mixture representations as input and conditions on the speaker embedding to estimate the mask for the desired speaker. The mask is multiplied on the mixture representations to generate representations for the target speaker. The the decoder transforms the target representations back to the time-domain signal. This time domain approach has been shown to improve the quality of the extracted signal compared with conventional T-F domain approaches [Xu et al., 2019b].

The best way to integrate the speaker embedding into the time-domain extractor network remains an open research question. Common strategies include multiplication (modulation)



(a) The speaker embedding is inserted to the extractor through multiplication. This approach has been used in Delcroix et al. [2020]

(b) The extractor is biased with the speaker embedding through concatenation. This approach has been used in Xu et al. [2019b]

Figure 5.3: Figures illustrating two contrasting approaches for introducing speaker embedding into the extraction network. a) multiplication, and b) concatenation.

and concatenation. The modulation operation is used in an end-to-end system known as time-domain SpeakerBeam (TD-SpeakerBeam [Delcroix et al., 2020]), illustrated in Figure 5.3a. The output of each unit after the first 1-D convolutional block in the extractor is modulated by the weights derived from the speaker embedding vector. The concatenation strategy has been used in a time-domain speaker extraction network (TseNet [Xu et al., 2019b]), shown in Figure 5.3b. The speaker embedding vector is repeatedly concatenated to the input features to each 1-D convolutional block.

5.2.3 Multi-channel speaker extraction

Multi-channel recordings have been exploited to improve the speaker extraction performance in noisy and reverberant environments. Popular methods taking advantages of multi-channel recordings in the speaker extraction task include the beamforming approach and exploiting spatial features. In fact, the first speaker extraction system in Žmolíková et al. [2017] was designed for multi-channel recordings. It combines the beamforming approach and the time-frequency domain neural network based mask estimator. Specifically, the estimated mask for the target speaker from a speaker extraction system is used for calculating coefficients for a generalised eigenvector beamformer (GEV).

Spatial features extracted from multi-channel signals have also been combined with spectral features to improve the speaker extraction's performance. Spatial features such as IPDs have proven to benefit time-domain extraction systems [Delcroix et al., 2020]. In Delcroix et al. [2020], spatial features were inserted into the middle layers after the speaker embeddings, as shown in Figure 5.4. Learned spatial features extracted by a 2-D convolutional layer can further improve the performance compared with the IPDs [Zorila et al., 2021]. However, the reported results show that these existing multi-channel extraction systems perform worse than multi-channel separation systems with similar architectures. It is suspected that existing feature fusion strategies for multi-channel speaker extraction cannot properly combine the spectral features, spatial features, and the speaker embeddings.

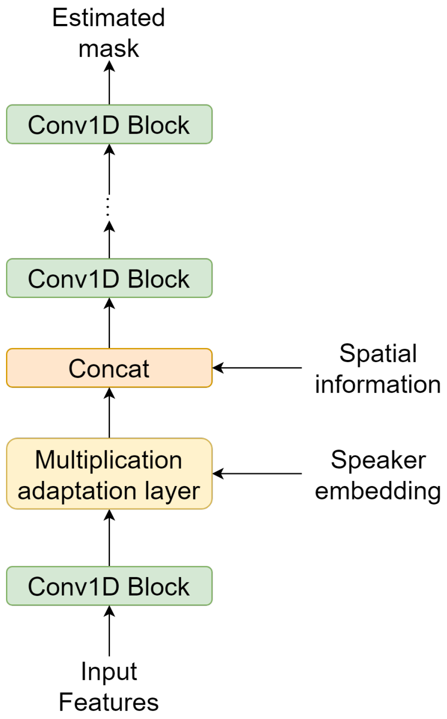


Figure 5.4: Time-domain SpeakerBeam with spatial features in Delcroix et al. [2020].

5.3 End-to-end multi-channel extraction

This section describes the design of the proposed time-domain multi-channel extraction system. The proposed system aims to effectively combine spectral features, spatial features, and speaker embeddings to improve speaker extraction performance. To achieve this, a novel speaker conditioning mechanism is introduced to the extraction system and will be described first in the following parts. Unlike the existing speaker extraction systems that target only one speaker each time, the proposed system considers simultaneously tracking multiple sources. In general, the system uses embeddings from multiple speakers as input in addition to the input mixture. These speaker embeddings are used to condition single-source outputs with a consistent speaker order. The second part of this section will consider the generation of speaker embedding and how to use multiple enrolled utterances to improve its robustness.

5.3.1 Network architecture

The proposed speech extraction network involves four components: an encoder, a speaker stack, a separation stack and a decoder. Compared with the separation network introduced in Chapter 4, the proposed structure incorporates a new speaker stack module. The overall structure is shown in Figure 5.5.

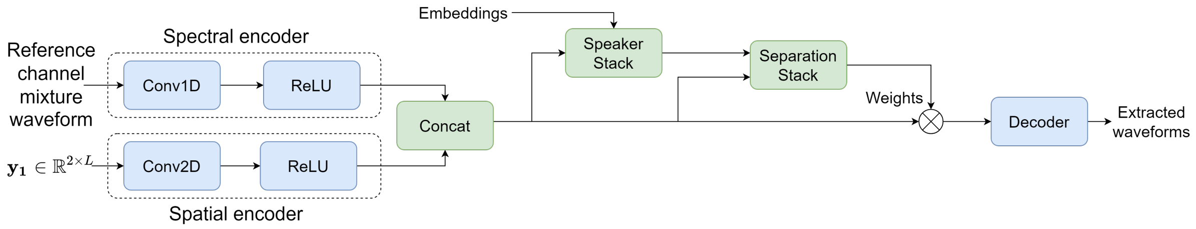


Figure 5.5: Proposed multi-speaker speech extractor with dedicated speaker stack

For the encoder, both single- and multi-channel cases are considered. For the single-channel extraction system, the encoder and decoder are constructed with 1-D convolutional layers. In a multi-channel speaker extraction system, the encoder consists of a spectral encoder and a spatial encoder. The spectral encoder is a 1-D convolutional layer followed with a ReLU activation function. The spatial encoder is constructed as a 2-D convolutional layer followed with a ReLU activation function. The concatenation of the encoded spectral features and spatial features forms multi-channel features, which are accessed by the other modules in the extraction system.

The question of how to supply identity information from target speakers to the separation stack is a key consideration in the design of a speaker extraction system. A recent speech separation system, Wavesplit, uses one convolutional subnetwork named as a speaker stack to map a mixture signal to a set of vectors representing speakers recorded in the mixture signal, which is then used to inform another convolutional subnetwork called a separation stack to estimate each source signal [Zeghidour and Grangier, 2021]. The separation stack is conditioned on the speaker representations with a Feature-wise Linear Modulation (FiLM) [Perez et al., 2018]. It was shown that the speaker representations reliably stabilise the separation for long speech sequences where dominant speakers can vary.

In this work, the architecture of the Wavesplit is adapted for the extraction system design. The designed speaker conditioning mechanism has a new 'speaker stack' to process the input speaker representations to coordinate with the main separation stack. The speaker stack takes the encoded features and generates two high-level sequential features, which are suitable for receiving speaker information from externally computed speaker embeddings (to be described in Section 5.3.2). The output of the speaker branch containing speaker information is then concatenated together with the encoded features as input to the separation stack. Note that the encoder is shared for both the speaker stack and the separation stack.

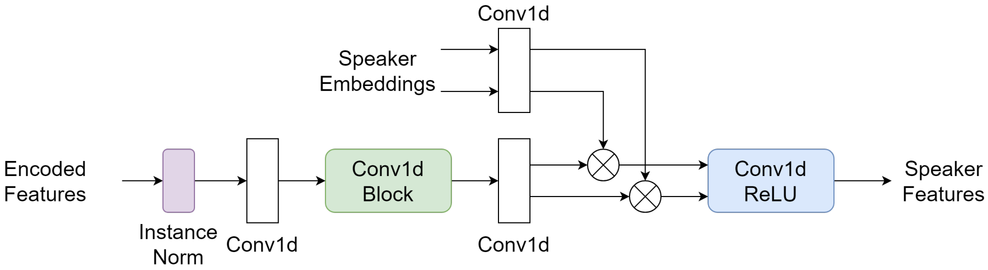


Figure 5.6: Internal structure of proposed speaker stack.

The speaker stack, illustrated in Figure 5.6, first employs an instance normalisation, a bottleneck 1-D CNN and a single TCN block to receive the encoded features. Then, the output of the TCN block will be factorised by an adaptation layer into multiple features for modulation with multiple speaker embeddings, which are transformed with a 1×1 convolutional layer to the same feature dimension. The modulated signals from each speaker embedding are concatenated together and processed with a 1-D convolutional layer and a ReLU non-linear activation function to form E-dimensional speaker information features, which have the same time length as the original features.

The separation stack is built by stacking multiple U-ConvBlocks. For a full description of the U-ConvBlock, the reader is referred back to Section 4.3. The separation stack takes the concatenation of the outputs from the encoder and the speaker stack as input and estimates multiple linear weights, each of which corresponds to a target speaker. The estimated linear weights are used to combine the encoded representations to generate separate representations for each speaker.

The speaker stack and the separation stack are jointly trained to directly optimise the SI-SNR metric. In contrast with PIT, the multi-speaker extraction system conditions the decoded signals on the speaker representations and keeps the output speaker order consistent with the order of input speaker embeddings. The proposed extraction system can also be used to perform single-speaker extraction. In that case, one speaker's embedding is input to the system and the extraction system outputs only one speech signal.

5.3.2 Increasing the robustness of speaker representations

A time-domain speaker recognition model, namely SincNet [Ravanelli and Bengio, 2018], is used for speaker embedding generation. Figure 5.7 illustrates the structure of the SincNet.

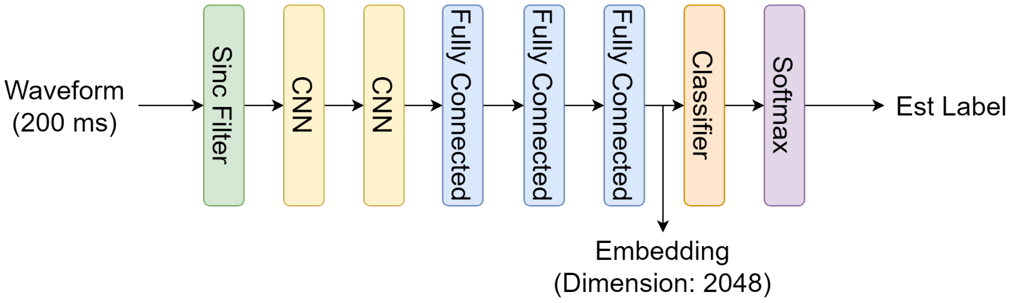


Figure 5.7: SincNet for speaker embedding generation

The SincNet employs a set of parameterised sinc functions as the first layer to encode a speech signal. This is equivalent to applying rectangular band-pass filters to a speech signal on the frequency domain. A band-pass filter $G(f)$ can be realised as the difference between two low-pass filters:

$$G(f, f_1, f_2) = rect(\frac{f}{2f_2}) - rect(\frac{f}{2f_1}) \tag{5.1}$$

where f_1 and f_2 denote the low and high cutoff frequency, respectively, and *rect* is the rectangular function. A band-pass filter on the frequency domain can be represented as a sinc function on the time domain $g(n)$:

$$g(n, f_1, f_2) = 2f_2 sinc(2\pi f_2 n) - 2f_1 sinc(2\pi f_1 n) \tag{5.2}$$

where the sinc function is defined as $sinc(x) = sin(x)/x$.

Following the sinc layers are two standard convolutional layers and four fully-connected layers. The last layer is a softmax classifier to estimate the speaker label. The input to the SincNet is a waveform segment of 200 ms. The SincNet is trained externally on a speaker recognition task with a cross-entropy loss.

Randomly selecting a single enrollment utterance for generating the speaker embeddings leads to unstable extraction performance [Li et al., 2019]. When multiple enrollment utterances from a target speaker are available, embeddings extracted from all utterances can be averaged to form a global embedding that represents a more stable speaker characteristics [Li et al., 2019]. Therefore, the same strategy is followed to obtain one global embedding for each speaker. When training the speech extraction system, one global speaker embedding is generated by averaging the utterance-level embeddings from all the training utterances belonging to the corresponding speaker. During evaluation, several utterances are randomly selected for each speaker, and the utterance-level embeddings from the selected utterances are averaged to form one global embedding.

Although increasing the number of used enrollment sentences can potentially improve the performance, long speech segments and multiple clean utterances from target speakers are hard to obtain. This trade-off problem will be investigated by comparing the performances with various available utterances for embedding generation when evaluating the extraction system.

5.4 Speech extraction experiment setup

The proposed methods are evaluated with a task that recovers anechoic speech for target speakers given a noisy and reverberant mixture. Both signal quality measurement (SI-SNR) and speech recognition performance (WER) are used as evaluation metrics. The following parts present the used data, configurations for the speech extraction network, configurations for the speaker embedding network, and setups for speech recognition evaluation.

5.4.1 Data

The evaluation of the proposed extraction methods is performed with simulated data, namely, the WHAMR! dataset [Maciejewski et al., 2020]. For a detailed description, the reader is referred back to Section 4.4. The extraction experiments use the training, validation, and test set partition defined in the official WHAMR! dataset. The speakers in the test set are unseen in the training and validation set, which forms an open-speaker-set evaluation.

The WSJ0 SI-84 corpus is used for training the speaker embedding network. The training set in WSJ0 SI-84 contains a total of 8769 unique utterances spoken from 101 speakers. The development and test set together contain 18 speakers. The speakers in the training set are different to the speakers in the development and test sets.

5.4.2 Speech extraction network

Separation models in both single- and multi-channel cases trained with PIT are set as the baseline for comparison. The baseline single-channel separation system uses the SuDoRM-RF model, which is constructed using U-ConvBlocks [Tzinis et al., 2020b]. The parameters of the SuDoRM-RF are set to match those of SuDoRM-RF 1.0x in Tzinis et al. [2020b], namely, $L = 17$, $B = 16$, $Q = 4$, $C = 256$, and $C_U = 512$. The baseline multi-channel separation system uses the single-stage multi-channel separation model constructed with U-ConvBlocks that was described in Chapter 4. The parameters are selected as $L = 17$, $S = 256$, $B = 16$, $Q = 4$, $C = 256$, and $C_U = 512$. The baseline multi-channel separation system achieved 12.8 dB SI-SNR improvement on WHAMR! dataset.

For the speech extraction system, the encoder, the separation model, and the decoder share the same configurations as the separation models. Several values for the dimension of the speaker features, E, in the speaker stack are evaluated, and $E = 128$ performs best empirically and is used to train the system. The input for all the models is the reverberated mixture with noise and the targets are the clean individual sources. The other training configurations are set to match those that were described in Chapter 4. To be specific, each utterance is split into multiple segments with a fixed length of four seconds. The Adam optimiser [Kingma and Ba, 2014] is used for training with a learning rate of 1e-3 and a batch size of four. The learning rate will be halved every time that the loss of the validation set is not reduced for three consecutive epochs.

5.4.3 Speaker embedding network

The SincNet [Ravanelli and Bengio, 2018] is trained on speakers in the training set of WSJ0 SI-84. The speakers in the development and test set are unseen during the SincNet training. This aims to investigate if the speaker embedding network and the speaker extraction network are able to generalise well to unseen speakers. The SincNet is trained on clean signals without noise and reverberation. During training, each speech signal is split into speech segments of 200 ms with 10 ms overlap. The SincNet employs the same configuration as in the original paper [Ravanelli and Bengio, 2018]. To obtain the speaker embedding used for the extraction task, the output of the last hidden layer of the final SincNet model is used to represent one frame-level speaker embedding for each 200 ms segment, and an utterance-level embedding is derived by averaging all the frame predictions. The dimension of the frame-level embedding and hence also the utterance-embedding is 2048.

5.4.4 Speech recognition evaluation setup

To evaluate the speech recognition performance, two acoustic models have been trained by accessing different conditions of training data. One model (AM1) was trained on the clean WSJ0/WSJ1 SI-284 data that contains roughly 80 hours data plus the WHAMR! single-speaker noisy reverberant speech that contains 116 hours data. In the official WHAMR! data, the single-speaker speech with noise is only simulated using the speech from the speaker with the higher SNR. However, here the speech with lower SNR is also used to simulate the single-speaker speech and this data is used during training the acoustic model in order to improve the noise robustness. The other one (AM2) was trained on the data used for AM1 plus the separated signals from the WHAMR! mixture in the training set processed by the proposed model. The audio data is downsampled to 8 kHz to match the sampling rate of data used for separation experiments. The acoustic model topology is a 12-layered Factorised TDNN [Povey et al., 2018], where each layer has 1024 units. The input to the acoustic model is 40-dimensional MFCCs and a 100-dimensional i-Vector.

A 3-gram language model is used during recognition. The acoustic model is implemented with the Kaldi speech recognition toolkit [Povey et al., 2011]. The ASR results obtained with AM1 on the standard clean WSJ Dev93 and Eval92 are 7.2% and 5.0% WER, respectively.

5.5 Experiments and results

This section presents the experiments performed to evaluate the proposed methods and analyses the results. Firstly, the proposed conditioning mechanism is compared to conventional strategies such as concatenation and multiplication. Next, we investigate the quality of speaker embedding that are generated with accessing various number of enrolled sentences and its impact on the speaker extraction task. Third, the multi-speaker extraction system is compared to the single-speaker extraction to investigate if the extraction of the target speaker can be improved by the knowledge of the identity of the other distracting speaker. Lastly, we analyse how the degree of speaker similarity in a mixture impacts on the extraction performance. The test set is split into three subsets conditioned on the pairs of gender in the two-speaker speech mixture (i.e., Female-Female, Male-Male or Female-Male). It is assumed that pairs with different genders are more distinct than pairs of speakers with the same gender.

5.5.1 Speaker conditioning mechanisms

This part conducts experiments to compare the proposed speaker conditioning mechanism with conventional approaches, including concatenation and multiplication. These conditioning approaches are evaluated for multi-speaker extraction. The systems are evaluated using the multi-channel recordings in order to show which mechanism can combine features extracted from multiple modalities more effectively. These features include spectral features, spatial features, and speaker identity features. The enrollment embedding is generated from one randomly selected utterance for each speaker.

The results are presented in Table 5.1. The baseline multi-channel separation model achieves

Table 5.1: Two-channel speech extraction performance with different speaker conditioning mechanisms. Concat and Multiply indicate the concatenation and multiplication conditioning mechanisms respectively. Split indicates the proposed conditioning mechanism.

Model	**PIT**	**SI-SNRi**
Multi-TasNet (U-Conv)	✓	12.8
Extraction (Concat)	✗	12.8
Extraction (Multiply)	✗	12.9
Extraction (Split)	✗	13.2
Extraction (Split)	✓	13.3

12.8 dB SI-SNR improvement without accessing speaker identity information. The extraction model with the conventional multiplication or concatenation strategies cannot directly benefit from the speaker information. They achieve similar results compared to the baseline separation system. The reason for failure of direct multiplication is presumed to be that the shared middle-stage features are not optimal for both tasks of speaker conditioning and speech reconstruction. As for the concatenation, the multi-channel features and the speaker embedding are completely different signals and cannot be suitably processed by the convolutional layer. It was pointed out that convolutional layers assumes time and frequency homogeneity of the input features [Wang et al., 2019b]. Conversely, the extraction model with the proposed mechanism can benefit from the speaker information and outperforms the blind source separation system and other conditioning strategies. The proposed method uses a separated speaker branch to generate high-level features for speaker conditioning tasks to alleviate the shared feature problem. And the sequential speaker features from the speaker branch can have a similar signal characteristic to the multi-channel features, which is a suitable input to the convolutional layers.

It should be noted that the proposed multi-speaker extraction system can be evaluated without accessing reference clean speech to find the right permutation. When the system is evaluated by accessing clean reference signals to find the oracle permutation, there is only a small difference between the two results. This demonstrates that our system can successfully identify and track multiple speakers in noisy and reverberant acoustic conditions.

Table 5.2: Average Duration (seconds) of Enrollment speech

#Utterance	1	2	3	6
Durations (sec)	6.6	12.8	20.9	41.6

Table 5.3: Speech extraction performance with speaker embedding generated from various number of utterances. Results shown for both single channel (nchs=1) and two channel systems (nchs=2). Performances measured in SI-SNRi (dB).

Model	**#nchs**	**#utters**	**SI-SNRi (dB)**
SuDo-RMRF [Tzinis et al., 2020b]	1	-	9.86
Extraction (split)	1	1	10.85
Extraction (split)	1	2	10.99
Extraction (split)	1	3	11.12
Extraction (split)	1	6	11.20
Multi-TasNet (U-Conv)	2	-	12.82
Extraction (split)	2	1	13.28
Extraction (split)	2	2	13.35
Extraction (split)	2	3	13.42
Extraction (split)	2	6	13.46

5.5.2 Influence of number of utterances for speaker embedding generation

This part investigates whether using multiple enrollment utterances can improve the speaker embedding and hence benefit speaker extraction. We select maximum six enrollment utterances for each speaker to generate speaker embedding. These sentences have been excluded from the data used for simulating mixtures for training and testing the speaker extraction system. Table 5.2 presents the average number of seconds of enrollment clean speech for embedding generation. In each condition, the duration is the averaged seconds over all test speakers.

The results of the baseline separation system and proposed speaker extraction systems are shown in Table 5.3. Since a separation system separates all speech signals from a mixture at the same time, the multi-speaker extraction system that also targets all speakers is used here to make a fair comparison. Both the separation and extraction systems access reference speech to find the right permutation. The SuDo-RMRF model is used as a single-channel

Table 5.4: Speech recognition performance with speaker embedding generated from various number of utterances. The two-channel extraction system is used to enhance the mixture. AM1 is used for decoding the extracted speech signals. Performances measured in WER (%).

Model	**#utters**	**SI-SNRi (dB)**	**WER (%)**
Mixture	-	-	78.49
Multi-TasNet (U-Conv)	-	12.82	34.86
Extraction (split)	1	13.28	33.50
Extraction (split)	2	13.35	33.24
Extraction (split)	3	13.42	32.79
Extraction (split)	6	13.46	32.77

baseline system without accessing speaker information, which achieves 9.86 dB of SI-SNR improvement. When the speaker embedding is extracted from a single enrollment sentence, the single-channel extraction system achieves 10.8 dB of SI-SNR improvement, which provides a large gain over the separation system. Increasing the number of utterances from one to three progressively improves the performance for both single- and multi-channel extraction systems. In the multi-channel case, using three utterances for embedding generation provides 0.14 dB gain of SI-SNR over using only one utterance, and the paired t test shows that the improvement is statistically significant ($t_{2999} = 2.39, P = 0.017$). This demonstrates that averaging embeddings from multiple utterances can improve the quality of the embedding, which leads to a performance improvement of a speech extraction system. However, further increasing the number of utterances beyond three yields limited SI-SNR gains.

The gain obtained from the speaker information for the multi-channel system is smaller compared with the single-channel system. The reason is possibly that the spatial information from multi-channel signals has already provided a strong cue for separation. The performance is further analysed in different conditions of angular separation between speakers in a mixture. Figure 5.8 shows that when the angle difference is small, the additional speaker identity information provides a consistent gain for both single- and multi-channel systems. When the angular difference is large, although the gain from the speaker identity information is reduced, it still provides the multi-channel extraction system with information complementary to the strong spatial information.

Table 5.4 presents the speech recognition results evaluated using AM1 that is trained on

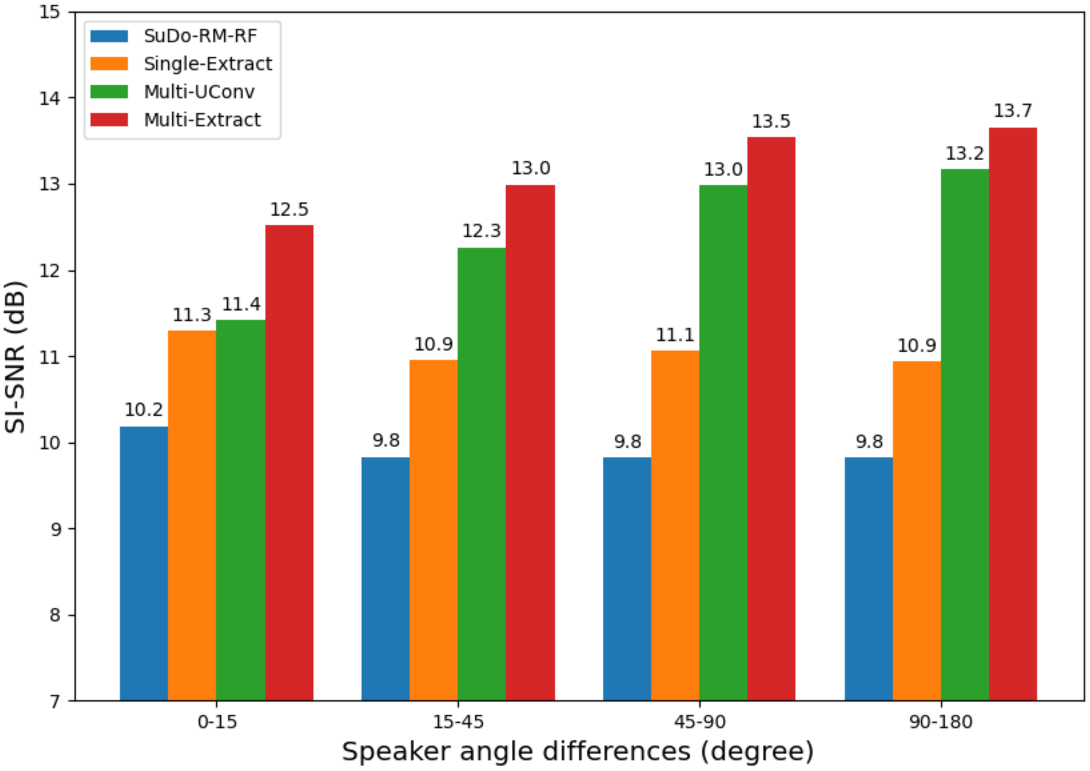


Figure 5.8: SI-SNRi performances of separation and extraction methods on different speaker angle conditions

clean WSJ data plus WHAMR! noisy and reverberant single-speech. The unprocessed mixture achieves 79.1% WER and the baseline separation system achieves 34.86% WER. By exploiting speaker identity information, the proposed speech extraction model yields a significant WER reduction over the noisy reverberant mixture and outperforms the strong multi-channel separation baseline. Further, improving the quality of speaker embedding by increasing the number of enrolled utterances yields slight improvements to the ASR performance. However, further increasing the number of utterances beyond three shows limited performance improvement. Therefore, in the following experiments, we select the same three utterances to generate the used speaker embedding for speaker extraction systems.

5.5.3 Comparison between single target and multiple targets

Although the proposed extraction system has been shown to be able to extract multiple speakers at the same time, it needs to be investigated whether an extraction system can benefit from

Table 5.5: Two-channel speech extraction performance comparison between single-target extraction and multi-speaker extraction.

Model	**#nchs**	**SI-SNRi**
SuDo-RMRF [Tzinis et al., 2020b]	1	9.9
Extraction (Two-target)	1	11.1
Extraction (Single-target)	1	11.5
Multi-TasNet (U-Conv)	2	12.8
Extraction (Two-target)	2	13.4
Extraction (Single-target)	2	13.9

multi-target training. Therefore, in this part, the multi-speaker extraction system is compared against a single-speaker extraction system. Given a mixture, a single-target system is trained to take embeddings from a single speaker to extract speech from this speaker each time. Both speakers in a mixture are used during training the single-target system. When evaluating the single-target system, both speakers in a mixture are extracted such that the performance can be compared directly with the multi-target approach.

Table 5.5 shows the results. The top part shows the performance for single-channel system and the bottom part shows the performance for extraction systems that access two-channel recordings. Interestingly, extraction systems targeting at a single speaker performs better than multi-target speaker extraction systems in both single- and multi-channel conditions. This observation suggests that extracting multiple speakers at the same time is harder than extracting a single speaker. The difficulty potentially comes from that the multi-speaker extraction system not only separates speech elements from a mixture signal, but also needs to associate these speech elements to different speakers, which could lead to speaker identity switches or lost tracks. However, a single-speaker extraction system does not face the speaker association problem. The similar object association issue has also been recognised for using neural networks to identify and track multiple objects from video recordings [Keuper et al., 2020; Zheng et al., 2021].

At this point, we would like to compare the proposed approach with other competing systems in the literature and the results are shown in Table 5.6. The proposed speaker conditioning mechanism provides a consistent separation performance gain in both single and multi-channel scenarios. With the additional information from multiple microphones and speaker enrollment,

Table 5.6: Comparative results of single and multi-channel speech separation/extraction on WHAMR! data

Model	**#nchs**	**Building Unit**	**PIT**	**SI-SNRi**
Conv-TasNet [Luo and Mesgarani, 2019]	1	TCN	✓	9.3
SuDo-RMRF [Tzinis et al., 2020b]	1	U-Conv	✓	9.9
Wavesplit [Zeghidour and Grangier, 2021]	1	TCN	✓	12.0
Nachmani et al. [2020]	1	RNN	✓	12.2
Multi-TasNet	2	U-Conv	✓	12.8
Extraction (Single-target)	1	U-Conv	✗	11.5
Extraction (Single-target)	2	U-Conv	✗	13.9

Table 5.7: Speech recognition results

System	#nchs	**WER(%)**	
		AM1	AM2
Mixture	-	78.5	77.0
Multi-UConv	2	34.9	-
Extraction (Two-target)	2	32.8	20.8
Extraction (Single-target)	2	30.2	20.2
Noisy Oracle	-	19.8	20.0

the proposed speaker extraction system achieves the best performance.

Table 5.7 compares the ASR results between single- and multi-target extraction systems. With the AM1 that is trained on clean WSJ data plus WHAMR! noisy and reverberant single-speech, the single-target extraction yields better recognition performance than the two-target extraction. These speech recognition results follow the same trend as the SI-SNR results achieved by these systems.

The time-domain approach usually introduces non-linear distortions to separated signals, which damage the speech intelligibility. An acoustic model is sensitive to the non-linear distortions if they are unseen during training, which is also known as a mismatch problem between training and evaluation. Therefore, another acoustic model, denoted as AM2, is trained with additional data augmented by the proposed extraction system to alleviate the mismatch problem. The fourth row in Table 5.7 shows that decoding the data with AM2, the WER is further reduced by 33% relatively, yielding an absolute 10.0% WER reduction compared to the result

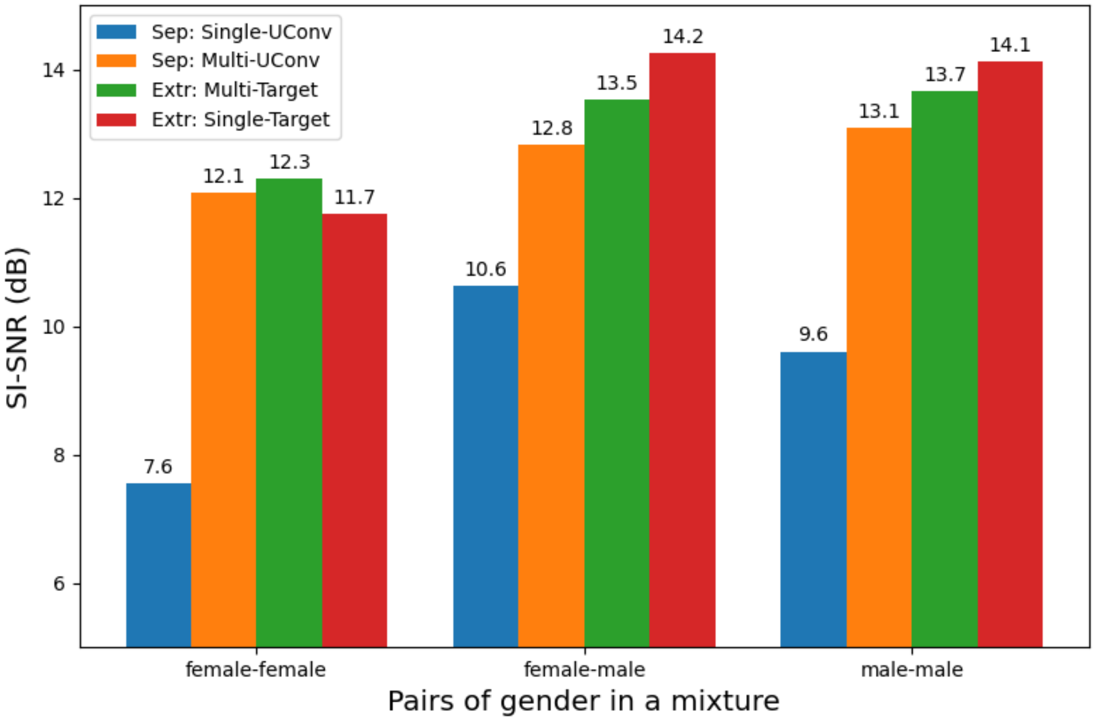


Figure 5.9: SI-SNRi performances of separation and extraction methods on different pairs of gender conditions, including female-female, male-female, and male-male.

achieved by AM1. The result achieved by AM2 is close to the result obtained with oracle single-speaker noisy reverberant speech (last row in Table 5.7).

5.5.4 Influence of speaker gender difference

It is hypothesised that when two speakers in a mixture are the same gender, it will be harder to separate than a mixture generated by two speakers of different genders. This part investigates how well the proposed systems perform with mixtures of same-gender speakers and different gender speakers in order to better understand how speaker similarities impact on separation process. The WHAMR! test set is split into three subsets, in each of which mixtures are generated in a unique pairs of gender condition, namely two female speakers, one male speaker and one female speaker, and two male speakers. The numbers of mixture samples for the female-female condition, the male-female condition, and the male-male condition are 394, 1520, and 1086, respectively.

Figure 5.9 compares the performances obtained by the single-channel separation system, the

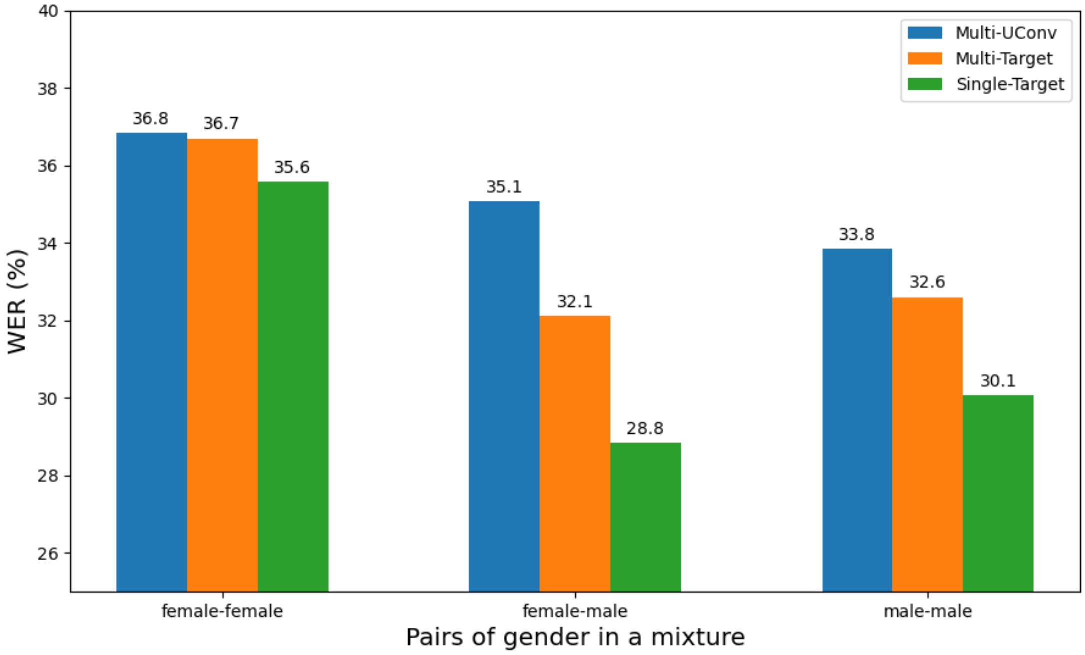


Figure 5.10: Speech recognition performances achieved by separation and extraction methods on different pairs of gender conditions, including female-female, male-female, and male-male.

multi-channel separation system, the multi-target extraction, and the single-target extraction system. The female-female condition is the most difficult case, in which all systems perform worse than the other two cases. For blind source separation, a single-channel system can achieve better separation performance with different gender mixtures than same gender mixtures. This result is foreseeable since the spectral characteristic between the two speakers in the different gender mixture is more different and easier to be separated than that in the same gender mixture. With the spatial information, a multi-channel system improves performance in all conditions and reduces the gap between the conditions.

By exploiting speaker identity information, the multi-target extraction system provides consistent gains over the separation system in all gender conditions. The single-target system further improves the performance over the multi-target system in both female-male and male-male conditions. However, in the female-female condition, the single-target system performs even worse than the separation system. This is potentially due to that, in some mixtures, two female speakers sound similar and the single-target extraction system treats both speakers as the same person. It suggests that when two speakers have a similar sound, identity information of the distractor could be helpful.

Figure 5.10 shows the recognition performances in the three mixture gender conditions. It can be observed that the WER in the female-female condition is higher than the other two conditions. This result is foreseeable because separating female-female mixtures is harder than the other two conditions and the separated signals of female-female mixtures have a lower speech intelligibility. The second observation is that the single-target extraction outperforms the multi-target extraction in all conditions. This is in contrast to the SI-SNR improvement results between multi-target extraction and single-target extraction in particular in the female-female condition. This suggests that the SI-SNR metric may not be optimal to develop front-end systems for the ASR .

5.6 Summary

In this chapter, we have presented a novel time-domain speech extraction system by extending the time-domain multi-channel separation system developed in the previous chapter. The design of this system has been improved from two aspects. The first aspect improved the quality of the speaker representations by averaging embeddings extracted from multiple enrolled utterances. Using the averaged embedding significantly improved the extraction performance compared with the embedding generated from a single enrolled utterance.

The second aspect considers the architecture design of the extraction system, which has introduced an additional speaker branch for receiving external speaker features. This mechanism solves the problems caused by combining features from multiple modalities, providing a more effective way for a multi-channel extraction system to use the speaker information to improve separation performance. Informed by multiple speaker embeddings, the proposed system is able to simultaneously output corresponding sources from a noisy and reverberant mixture, without a label permutation ambiguity. The extraction system equipped with the proposed conditioning mechanism yields better speech separation result compared to extraction systems with conventional conditioning methods such as multiplication and concatenation.

It is also found that the multi-target extraction system performs slightly worse than the system extracting a single target each time. It is possibly due to that the task of separating multiple speakers is more difficult than the task of targeting one speaker. When applied as front-end processing for ASR, the single-target extraction system yields better result than the multi-target extraction system.

End-to-end speaker extraction systems usually introduce non-linear distortions, which hurt speech recognition performance. This work reduces this impact by training an acoustic model using additional data processed by the end-to-end extraction system, therefore the condition of the ASR training data matches the condition of the extracted signal. The speech recognition performance is improved from 30.2% WER achieved by the mismatched acoustic model to 20.2% WER achieved by the matched acoustic model.

Further analysis is conducted to show how speaker similarities impact the separation and recognition performance. It is found that mixtures from two female speakers are harder to separate compared with mixtures from different genders or two male speakers. The separation performance of single-target extraction is slightly worse than multi-target extraction with female-female mixtures. Conversely, for overlapping speech recognition, the single-target extraction outperforms the multi-target extraction system in the female-female mixture condition. These contradictory results suggest that the SI-SNR metric may not be optimal to develop front-end methods for ASR.

Current deep learning based separation and extraction systems require large amounts of paired mixture and reference signals to conduct supervised training to achieve success. This limits the applicability of these methods to real scenarios because matched noisy mixtures and clean reference targets are hard to collect in practice. In next Chapter, we will study how to build a separation network in situations where only a limited amount of clean reference signals are accessible.

Chapter 6

Unsupervised and semi-supervised end-to-end speech separation

6.1 Introduction

Most current time-domain separation networks are developed using supervised training, which requires large amounts of paired noisy mixtures and target reference signals. This paired data is usually obtained by using simulation tools that simulate noisy mixtures by mixing multiple single-speaker speech and noise signals [Hershey et al., 2016; Maciejewski et al., 2020; Wang et al., 2018]. However, in real scenarios, although large amounts of noisy data may be readily available, paired clean reference signals are hard to collect.

Although separation models trained with simulated data have achieved impressive performances when evaluated with matched simulated data, they often perform poorly on real mixtures. This is due to unavoidable mismatches between simulated and real environments. This is because the distribution of sound types and acoustic conditions of real environments are hard to measure and the current simulation tools cannot simulate data fully matched to the real recordings. In addition, current simulation algorithms have not fully considered other factors that could affect acoustic conditions, including speaker movements, directivity patterns for human speech, reflections from objects except walls, etc. The mismatch problem between simulated and real conditions cannot be fixed simply by retraining with 'real data' because in real scenarios there is generally no access to isolated ground truth signals.

To exploit only unlabelled noisy signals to build speech separation systems, efforts have been undertaken to design unsupervised learning approaches [Drude et al., 2019a; Wisdom et al., 2020]. An unsupervised algorithm, known as mixture invariant training (MixIT) [Wisdom et al., 2020], has been proposed as a solution to training separation models without the requirement of isolated ground truth signals. The MixIT approach is described in detail in Section 6.2 but the basic idea is that it only uses mixture signals during training. A mixture of mixtures is input to the model and the estimates of individual sources are recombined to reconstruct reference mixtures. The estimated individual sources can be recombined in many permutations, i.e., arbitrary assignment of components to mixtures, and the estimated mixtures are based on the permutation that best reconstructs the reference mixtures. A signal construction loss is computed between the original mixtures and the estimated mixtures. During inference, since a MixIT model takes a single mixture signal as input, it usually produces more signals than sources. The estimated speech signals should be selected from the model outputs and this selection process might cause some errors. Furthermore, since the model produces a greater number of estimates than the sources present in a single mixture, it causes an over-separation problem [Wisdom et al., 2021, 2020]. This leads to poor separation in the sense of splitting a single source across multiple outputs.

An unsupervised separation model can be significantly improved via fine-tuning when a small amount of supervised training data is available [Wisdom et al., 2020]. Unsupervised training uses unlabelled data in a task-agnostic fashion, while fine-tuning can exploit a few supervised examples to adapt the pre-trained model for a specific task. A recent work in computer vision has shown that a bigger model can benefit more from the task-agnostic use of unlabelled data, while a task-specific model may be best structured in a different way and have a smaller model size [Chen et al., 2020a]. So, once a model is pre-trained and fine-tuned, the knowledge of this model can be *distilled* to a compact student model for further improvement. The student model is trained to mimic the predictions generated by the fine-tuned model processing the unlabelled data. However, the current semi-supervised framework for speech separation system has not considered exploring different model architectures.

This chapter makes two contributions. We first develop a fully unsupervised approach based on MixIT but incorporating solutions to the over-separation problem. This is then followed by explorations of semi-supervised approaches that attempt to match the performance of the previous fully supervised approaches but with a fraction of the data.

The first piece of work develops a novel unsupervised algorithm to address the over-separation problem of the MixIT. It is hypothesised that reducing the number of output channels of the separation model and training this model with single mixtures can alleviate this problem. Based on this assumption, the designed algorithm exploits teacher-student learning to transfer the knowledge from a MixIT model to a student model with a smaller number of output channels. A teacher model trained with MixIT processes single mixtures to generate multiple outputs, from which speech signals are selected with an energy restriction. The selected speech signals are treated as labels for the single mixture and used to train the student model in a standard supervised fashion [Kolbæk et al., 2017].

The second piece of work provides a solution to applying end-to-end separation techniques to situations where there are a large number of mixture signals and just a few reference source signals. This solution is a semi-supervised approach comprising three steps: 1) unsupervised training with mixtures, 2) supervised fine-tuning with supervised training data, 3) knowledge distillation using mixtures. For the unsupervised training step, the proposed unsupervised algorithm is used. With the small amount of supervised training data, the unsupervised model is fine-tuned to be more task-specific. Finally, a second knowledge distillation stage is used to generate a more compact task-specific model, increasing performance and efficiency in the process.

The rest of this chapter is structured as follows. Section 6.2 describes the background. In Section 6.3, the proposed unsupervised and semi-supervised approaches are described. Implementation details and experiment setup are described in Section 6.4. Section 6.5 presents the experiments and analyses the results.

6.2 Background

This section describes an existing supervised learning framework and an unsupervised learning framework that will be used to build the proposed methods in Section 6.3. In the supervised learning case, a mixture x containing up to P sources and its corresponding ground truth signals $\mathbf{s}$ are given. A separation model takes the mixture waveform x as input and predicts $M = P$

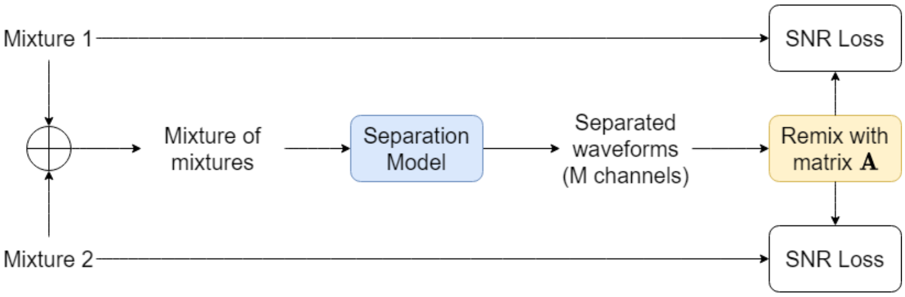


Figure 6.1: MixIT training algorithm

sources. Permutation invariant training (PIT) [Kolbæk et al., 2017] has become a standard training framework for training a separation network in a supervised way. Since the order of the predicted signals is arbitrary, $P!$ permutations of prediction to ground truth pairs exist for computing the signal reconstruction loss. PIT addresses this problem by using the permutation which gives the smallest loss computed over the entire utterance. The smallest loss will be used to back-propagate to optimise the model parameters.

In the unsupervised learning case, a speech separation system is trained only on unlabelled signals. An unsupervised learning approach, namely, mixture invariant training criterion (MixIT), has been recently developed to train a separation system in situations where the ground truth signals are not available [Wisdom et al., 2020]. The MixIT training framework is illustrated in Figure 6.1. The MixIT draws two mixtures x_1 and x_2 at random without replacement from an unsupervised training dataset. Each mixture x_i contains up to P_i sources. The two selected mixtures form a mixture of mixtures by adding them together: $\overline{x} = x_1 + x_2$. An end-to-end separation model takes $\overline{x}$ as input and predicts $M \geq P_1 + P_2$ source signals, which will be used to reconstruct the two mixtures. The unsupervised MixIT loss is computed between the estimated sources $\hat{\mathbf{s}}$ and the input mixtures x_1, x_2 as follows:

$$\mathcal{L}_{\text{MixIT}}(x_1, x_2, \hat{\mathbf{s}}) = \min_{\mathbf{A}} \sum_{i=1}^{2} \mathcal{L}(x_i, [\mathbf{A}\hat{\mathbf{s}}]_i), \tag{6.1}$$

where $\mathbf{A} \in \mathbb{B}^{2\times M}$ is a mixing matrix whose elements along each column sum to 1. The mixing matrix assigns each $\hat{s}$ to either x_1 or x_2. The estimated sources $\hat{s}$ are the time-domain outputs from the decoder in the separation model.

The signal-level loss function between a reference $y \in \mathbb{R}^T$ and estimate $\hat{y} \in \mathbb{R}^T$ from the separation model is the negative threshold SNR:

$$\begin{aligned}\mathcal{L}(y, \hat{y}) &= -10\log_{10}\frac{\|y\|^2}{\|y-\hat{y}\|^2 + \tau\|y\|^2} \\ &= 10\log_{10}(\|y-\hat{y}\|^2 + \tau\|y\|^2) - 10\log_{10}\|y\|^2,\end{aligned} \tag{6.2}$$

where $\tau = 10^{-\mathrm{SNR}_{max}/10}$ acts as a soft threshold that clamps the loss at SNR_{max}= 30 dB, which is the same value as in Wisdom et al. [2020]. T is the signal's length in samples.

As described in the original paper, a mixture consistency constraint [Wisdom et al., 2019] has been introduced during training the MixIT model, which forces the network to predict sources that sum up to the mixture. This constraint solves the following optimization problem to find mixture consistency separated sources $\hat{s}_m$ given initial separated sources $\underline{s}_m$:

$$\min_{\hat{s}\in\mathbb{R}^{M\times T}} \frac{1}{2}\sum_m \|\hat{s}_m - \underline{s}_m\|^2 \text{ subject to} \sum_m \hat{s}_m = x. \tag{6.3}$$

The closed-from solution of this problem is:

$$\hat{s}_m = \underline{s}_m + \frac{1}{M}(x - \sum_{m'} \underline{s}_{m'}). \tag{6.4}$$

In Wisdom et al. [2020], the MixiT framework has been evaluated for a single-channel separation task in both anechoic and reverberant conditions. During evaluation, the trained MixIT model is used to process a mixture of two speakers to output M signals. To compute the SI-SNR score, reference sources are zero-padded to M signals as well, which are aligned to the model outputs with a permutation that maximises the SI-SNR score. Then the SI-SNR scores for non-zero reference signals are averaged to achieve the final SI-SNR score for one mixture sample. It has been shown that the unsupervised MixIT framework achieves a performance that is worse by up to 7 dB compared to a fully supervised system in the anechoic condition, and achieves a performance that is worse by 3 dB compared to a fully supervised model in the reverberant condition. In addition, it was observed that, with a small amount of unsupervised data from a matched domain, the MixIT framework can effectively reduce the domain mismatch problem caused by training a separation model with mismatched supervised data.

Although the MixIT framework has made a great contribution to unsupervised learning for speech separation networks, it suffers from the mismatch between training and test and a tendency to 'over-separate' the mixture. Firstly, since the number of speakers in an input mixture of mixtures during training is always larger than that from a single mixture during inference, there is a severe mismatch between training and evaluation and this mismatch can cause the MixIT model a severe over-separation issue. The over-separation issue means that the separation model could split a single source across multiple outputs.

6.3 Proposed method

In order to solve the mismatch and over-separation problems, we wish to use single mixtures as input during training and reduce the number of output channels of the separation model to match the number of sources in the mixture signals. Based on this, this section first proposes an unsupervised training algorithm that combines both MixIT and PIT criterion via teacher-student learning. Followed by the proposed unsupervised learning approach, this section next introduces a novel semi-supervised approach that attempts to applying speech separation techniques to situations where a small number of supervised training data is available.

6.3.1 Teacher-student MixIT

The teacher-student training process for mixtures of two speakers is illustrated in Figure 6.2. We first train a teacher model on mixture of mixtures using the MixIT criterion. Then, the teacher model is used to process the original mixtures to generate M isolated signals. Since the number of teacher model outputs is larger than the number of sources in a mixture, target speech signals should be selected from all the outputs. Here we consider to use an energy restriction, but other selection criteria that are related to speech and speaker properties might be applied in the future. The energy restriction is to select output streams with the highest energies, based on the assumption that the number of speakers in the mixture P is known and the target speakers will correspond to the streams with the highest energies. Therefore, P output streams with the highest energies will be selected as the pseudo-targets. Then, the student model is trained to separate the same mixture sample into the selected separated sources with the PIT framework. The negative threshold SNR described in Equation 6.2 is used as the signal reconstruction loss function during training the student model.

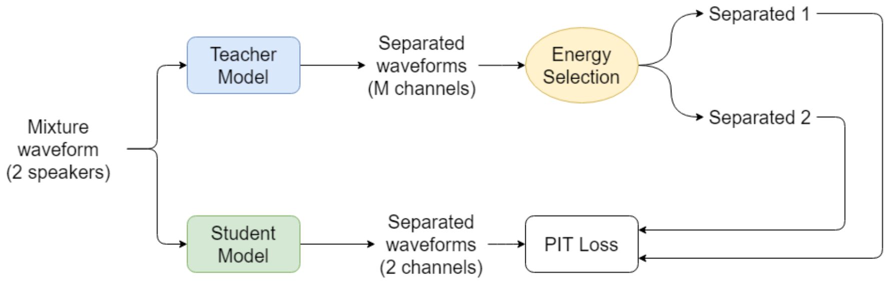


Figure 6.2: Teacher-student MixIT training algorithm

6.3.2 Fine-tuning

When a few supervised training data is accessible, we would like to exploit both large amounts of unsupervised training data and the small amount of supervised training data to improve the separation model. Training a model using just a very small amount of supervised training data from scratch usually causes an overfitting problem, meaning that the trained model can not generalise well to unseen situations. However, if the model has been initialised via unsupervised training with a large amounts of unlabelled data before supervised training, the generalisation of the model can be improved [Erhan et al., 2010].

The proposed teacher-student MixIT model can be considered as an unsupervised pre-trained model. Unsupervised pre-training teaches the model to learn a general representation that can be transferred to downstream tasks [Devlin et al., 2018; Oord et al., 2018; Schneider et al., 2019]. When a small number of parallel mixture and target signals are available, the unsupervised trained model can be fine-tuned to be more task-specific. This work uses a simple strategy for fine-tuning which trains the entire pre-trained student model without replacing or discarding any layers. The network is trained in the standard PIT framework with the negative SNR loss. There are more complex strategies for model fine-tuning or transfer learning in the literature, e.g., freezing or replacing layers, and adjusting the learning rate [Tzinis et al., 2020a; Zhang et al., 2020b], which can be explored in the future work.

6.3.3 Knowledge distillation via unlabeled examples

It has been shown in an image classification task that deeper and bigger models can benefit more from the task-agnostic use of unlabeled data, while a task-specific model may have a different

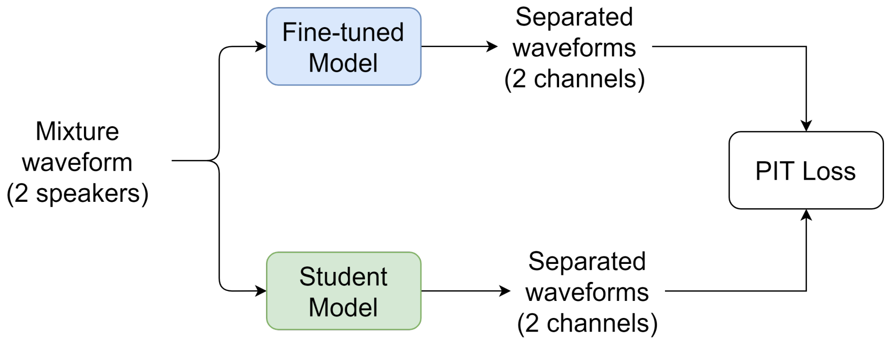

Fine-tuned Model
Separated waveforms (2 channels)
Mixture waveform (2 speakers)
PIT Loss
Student Model
Separated waveforms (2 channels)

Figure 6.3: Knowledge distillation diagram. The fine-tuned model serves as a teacher model to train another compact separation model.

structure, which improves the task-specific performance, or have a smaller model architecture, which leads to a compact model [Chen et al., 2020a]. The knowledge from the big unsupervised model can be exploited by a small task-specific model to further improve performance. With this observation as motivation, we would like to employ different network architectures during the semi-supervised learning process for the speech separation task.

Figure 6.3 illustrates the knowledge distillation process for the semi-supervised speech separation. A big model is used during the unsupervised learning and fine-tuning. Then, the fine-tuned model is used as a teacher to generate pseudo-targets again for training a student network with a different network architecture. Any compact model that achieved better performance with supervised training could serve as the student in the knowledge distillation stage. During the knowledge distillation, the teacher network is fixed and only the parameters of the student network are optimised. Again, the student model is trained in the standard PIT framework with the negative SNR loss.

6.4 Experiment setup

Since the main concern in this thesis is overlapping speech recognition, the separated speech from proposed separation systems is decoded by an ASR system and the recognition performance (WERs) is evaluated. Signal quality measurement (SI-SNR) is also reported to analyse the separation performance.

The experiments are conducted using simulated data, considering two acoustic conditions.

One is an anechoic condition without noise and the other is a noisy and reverberant condition. The proposed training framework is applied to both single- and multi-channel separation networks. The following parts will described the used data, separation network configurations and speech recognition setups.

6.4.1 Data

The WSJ0-2mix dataset [Hershey et al., 2016] is used for experiments in the anechoic condition. The WHAMR! dataset [Maciejewski et al., 2020] is used for the experiments in the noisy and reverberant condition. It uses WSJ0 SI-84 data to simulate noisy and reverberant 2-speaker mixtures with a stereo (2-channels) configuration. The room impulse responses are artificially generated using the Pyroomacoustics toolkit [Scheibler et al., 2018]. For details of WHAMR! data simulation configurations, the reader is referred back to Section 4.4. For both the anechoic and the noisy case, there are 20k sentences from 101 speakers for training, and 3k sentences from 18 speakers for testing. The speakers in the test set do not appear during the training of the speech separation system. All data has been downsampled from a 16 kHz sampling rate to an 8 kHz sampling rate.

During training the MixIT model, the mixture of mixture are dynamically generated from randomly selected mixture signals in the official training set. A constraint is set to guarantee that the four speech signals in every mixture of mixtures come from different speakers. The first 2k sentences in the official training set are selected for supervised training and fine-tuning, which accounts for 10% of the total official training data.

To investigate how the mismatch problem caused by the artificial training mixtures affects a MixIT model, the mixture of mixtures are created with two strategies: i) Both mixtures always consist of two speakers (*2-src*) such that the mixture of mixtures always contain four sources. ii) One of the mixtures use either a single or two speakers (*1or2-src*) such that the mixture of mixtures contain three or four sources. In the later case, the single-talker signal comes from the selected 10% of training data for fine-tuning.

6.4.2 Separation network configurations

For the single-channel separation task, the Conv-TasNet [Luo and Mesgarani, 2019] architecture is employed as the unsupervised model. Conv-TasNet's hyper-parameters are set as those that produced best performance in the original paper namely, $N = 256$, $B = 128$, $H = 256$, $L = 20$, $R = 4$, $X = 7$. The number of output channels M for the teacher model and the student model in the anechoic condition is set as four and two, respectively. For the multi-channel separation task, the TCN based multi-channel separation network (Multi-TCN) described in Chapter 4 is employed as the unsupervised model. The hyper-parameters are set as those used in Chapter 4: $N = 256$, $S = 36$, $R = 3$, $X = 7$, $L = 16$. The number of output channels M for the teacher model and the student model in the noisy and reverberant condition is set as eight and two, respectively. In the student model, the application of the mixture consistency constraint depends on the task. It is employed when the separation is under the anechoic condition since the separated sources should sum up to the original mixture. However, in the noisy and reverberant condition, it is removed since the student model only outputs speech signals and discards noise signals. During training the separation model, the input is a segment with a fixed length of four seconds which is obtained by splitting each utterance. All models are trained with the Adam optimiser [Kingma and Ba, 2014] with a learning rate of 1e-3 and a batch size of eight.

For knowledge distillation, we used separation models outperforming the Conv-TasNet and the Multi-TCN in the supervised learning framework, while maintaining a smaller model size. The dual-path RNN (DPRNN) [Luo et al., 2020] and the U-Convolutional block (U-ConvBlock) based multi-channel separation network have been selected for the single- and the multi-channel tasks, respectively. The hyper-parameters of DPRNN are set as those that produced the best performance in Luo et al. [2020]. The hyper-parameters in the U-ConvBlock multi-channel separation network reuse those in Chapter 4 as follows: $N = 256$, $L = 17$, $S = 256$, $B = 16$, $Q = 4$, $C = 256$, and $C_U = 512$.

6.4.3 Speech recognition evaluation

To evaluate the speech recognition performance, two acoustic models, AM1 and AM2, have been trained using the Kaldi speech recognition toolkit [Povey et al., 2011]. The model AM1 was trained on roughly 80 hrs of clean WSJ0/WSJ1 SI-284 data plus the WHAMR! single-speaker noisy reverberant speech. In the official WHAMR! data, the single-speaker speech with noise is only simulated using the speech from the speaker with the higher SNR. However, here the speech with lower SNR is also used to simulate the single-speaker speech and this data is used during training the acoustic model in order to improve the noise robustness. The second model, AM2, was trained on the data used for AM1 plus the separated signals from the training mixtures of WHAMR! processed by the proposed multi-channel model with fine-tuning and knowledge distillation. The audio is downsampled to 8 kHz to match the sampling rate of the data used for the separation experiments. The acoustic model topology is a 12-layer factorised TDNN [Povey et al., 2018], where each layer has 1024 units, and it is trained using 40-dimensional MFCCs and 100-dimensional i-Vectors. Recognition is performed using a 3-gram language model provided in the Kaldi baseline WSJ recipe. With our set-up, the word error rate (WER) results obtained with AM1 on the standard clean WSJ Dev93 and Eval92 are 7.2% and 5.0%, respectively.

6.5 Experiments and results

This section presents experiments and results to evaluate the proposed unsupervised and semi-supervised approaches. The first part evaluates the proposed teacher-student MixIT method in two different environmental conditions. The second part investigates the semi-supervised approaches, including model fine-tuning and knowledge distillation. The scale-invariant signal-to-noise ratio (SI-SNR) is used to measure the separation performance, and the WER is used to measure the speech recognition performance. These scores are computed using both of the two estimated speech signals obtained from a separation model.

6.5.1 Unsupervised speech separation

Anechoic condition

We first conduct experiments in the anechoic condition to investigate the over-separation issue in the MixIT approach. To obtain estimated speech from the MixIT model outputs, two strategies are considered. The first is based on an energy criterion such that the two output signals with

Table 6.1: Results for anechoic single-channel separation using unsupervised learning (Conv-TasNet architecture)

System	**Mixes**	**M**	**SI-SNRi (dB)**
MixIT (Energy)	2-src	4	9.0
MixIT (Oracle)	2-src	4	9.8
TS-MixIT	2-src	2	10.4
MixIT (Energy)	1or2-src	4	12.0
TS-MixIT	1or2-src	2	12.6
Supervised (10%)	-	2	11.3

the highest signal energies are selected from the M output channels. Another strategy is to remix the output channels to generate two estimated signals. In this experiment, the model outputs are remixed in all permutations and the one that achieves the highest SI-SNR score is selected. The result achieved by this strategy can be seen as an oracle result, representing the best performance that could be achieved if the over-separation problem was solved.

The performances achieved by the MixIT models are presented in the first two rows of Table 6.1. The oracle remixing achieves 0.8 dB improvement over the energy selection mechanism. Since the mixture during test contains only two speech signals, the outputs of a separation model are expected to contain two non-zero signals and zero signals for the other. The performance drop of the energy selection compared to the remixing strategy comes from the over-separation problem of the MixIT model. When some mixture of the mixtures are created with 1or2-source signals, the MixIT model achieves a 3.0 dB gain of SI-SNR over the model trained with only two source mixtures, comparing the results in the first and fourth rows of Table 6.1. Accessing single-speaker speech signal as training targets can be considered as supervised training, which reduces the effect of over-separation.

Next, we investigate if using teacher-student learning can alleviate the over-separation issue. As shown in Table 6.1, the proposed student model outperforms the MixIT model with energy selection. Since the final student model has a smaller number of output channels than the MixIT model, this result suggests that TS-MixIT effectively addresses the over-separation issue, therefore the target signals in the student model can be selected more easily and accurately. The proposed student model even outperforms the MixIT model with oracle selection. Furthermore, when the MixIT model is trained with 1or2-source mixtures, the proposed student model still provides 0.6 dB gain over the MixIT model, as shown in the fourth and fifth rows in Table 6.1. This is because the input mixture to the student model during training has a matched condition as the input mixture used in the test. However, a mixture of mixtures for training the MixIT model usually contains more speakers than that in a mixture for test, which causes a mismatch problem. Compared to the supervised model that is trained with the 10% of the training data (the final row in Table 6.1), both the MixIT and TS-MixIT models that are trained on 1or2-src mixtures achieve better separation performances. This indicates that the unsupervised approaches can exploit the unsupervised training data to improve the generalisation of the model.

Next, we ask whether the unsupervised speech separation system can benefit ASR with overlapping speech. The experiment uses the clean WSJ0/WSJ1 SI-284 for acoustic model training. Although the anechoic condition is unrealistic, we use it for theoretical analysis. The separated speech signals from the separation model are decoded by the pre-trained acoustic model.

Table 6.2 presents the speech recognition accuracy. The results in the top part shows that the separation models trained in both MixIT and the proposed TS-MixIT frameworks provides significant WER reduction compared with the original unprocessed mixture. The proposed system (TS-MixIT trained with the 2-src mixture) yields 7% absolute WER reduction compared with the baseline system (MixIT trained with the 2-src mixture). The result suggests that the over-separation problem in the MixIT model has a negative impact on the ASR performance and the proposed unsupervised learning approach can effectively address this issue.

The middle part shows the results when the MixIT model is trained with the 1or2-source mixtures. In this case, both MixIT and TS-MixIT models achieve better ASR results compared to model trained with the 2-source mixture. Accessing single source signal as target improves the quality of the separated signals, hence leading to better recognition performance. The proposed student model provides 0.5% WER reduction compared to the MixIT model (i.e. a reduction of 26.7% to 26.2%). Since the over-separation problem of the MixIT model is alleviated by accessing single source signals, the benefit from the proposed method is narrowed.

Table 6.2: Speech recognition performance of unsupervised separation under the anechoic condition without noise and reverberation

System	**WER(%)**
Mixture	79.6
MixIT (2-src)	37.5
TS-MixIT (2-src)	30.6
MixIT (1or2-src)	26.7
TS-MixIT (1or2-src)	26.2
Supervised (100%)	19.3
Oracle	9.1

Noisy and reverberant condition

This part investigates if TS-MixIT can improve the separation performance in more realistic conditions including both reverberation and additional non-speech noise sources. A single-channel separation model is trained with MixIT as a baseline model. Then multi-channel separation systems are trained with both MixIT and TS-MixIT. The estimated speech signals are selected from the MixIT model based on the energy criterion. To generate 1or2-source mixtures, a reverberant single source signal is mixed with a noisy reverberant mixture. The SI-SNR score is computed between the estimated signals and reverberant single speech references.

Table 6.3 shows the SI-SNR gain of the unsupervised separation networks. The first observation is that the performance of the single-channel MixIT model drops significantly from 9.0 dB in the anechoic condition to 3.5 dB in the noisy condition. This demonstrates the difficulty of unsupervised separation with noise and reverberation. Secondly, the results show that the spatial information helps improve the separation accuracy of models trained with the MixIT criterion. By accessing two-channel recordings, the MixIT model improves separation performance by 2 dB (comparing row 1 and row 2). Furthermore, TS-MixIT outperforms the baseline MixIT model for both data mixing approaches. The TS-MixIT approach reduces the number of output channels in the separation model from eight to two, which effectively alleviates the selection and over-separation problems.

To understand how spatial information affects the unsupervised learning approach, the separation performance is measured in conditions with various speaker angular differences. We

Table 6.3: Results for 2-channel denoising and separation using unsupervised learning (Conv-TasNet architecture)

System	Mixes	#nchs	M	SI-SNRi (dB)
MixIT (Energy)	2-src	1	8	3.5
MixIT (Energy)	2-src	2	8	5.6
TS-MixIT	2-src	2	2	6.5
MixIT (Energy)	1or2-src	2	8	6.1
TS-MixIT	1or2-src	2	2	7.7
Supervised (10%)	-	2	2	8.5

compare the performances achieved by the single- and two-channel MixIT models, and the two-channel TS-MixIT models. Figure 6.4 shows that, when the angular difference increases, the multi-channel systems gradually improve the separation performance by taking advantages of the increasing spatial difference between the two sources. It indicates that the unsupervised approach can force the multi-channel separation model to exploit the spatial information without explicit supervision. When the two sources are spatially separable with a angular difference larger than 15 degrees, the multi-channel system benefits from the spatial information and outperforms the single-channel system. However, when the angular difference is small, the single-channel MixIT model outperforms the two-channel MixIT model. It suggests that the spatial information has a negative effect on the separation system when the speakers are not spatially separable.

This part investigates how speaker similarities impact the unsupervised approaches in the noisy condition. The test set is split into three subsets specified by the mixture gender conditions (i.e., male-male, female-female and male-female). The performances of MixIT and TS-MixIT models are compared in different subsets. The results are shown in Figure 6.5. Firstly, we see that it is difficult for the single-channel MixiT model to separate mixtures under the same gender mixture condition. The single-channel system achieves 2.3 dB and 1.6 dB SI-SNR gain in the female-female pair and the male-male pair, respectively. Secondly, the Multi-MixIT provides significant gains over the single-channel MixIT model in the same gender mixture. This demonstrates that the unsupervised model exploits the spatial cues for separation. However, the benefit from the spatial information is limited in the condition of different gender mixture. It seems that the network can only rely on one type of information for separating one

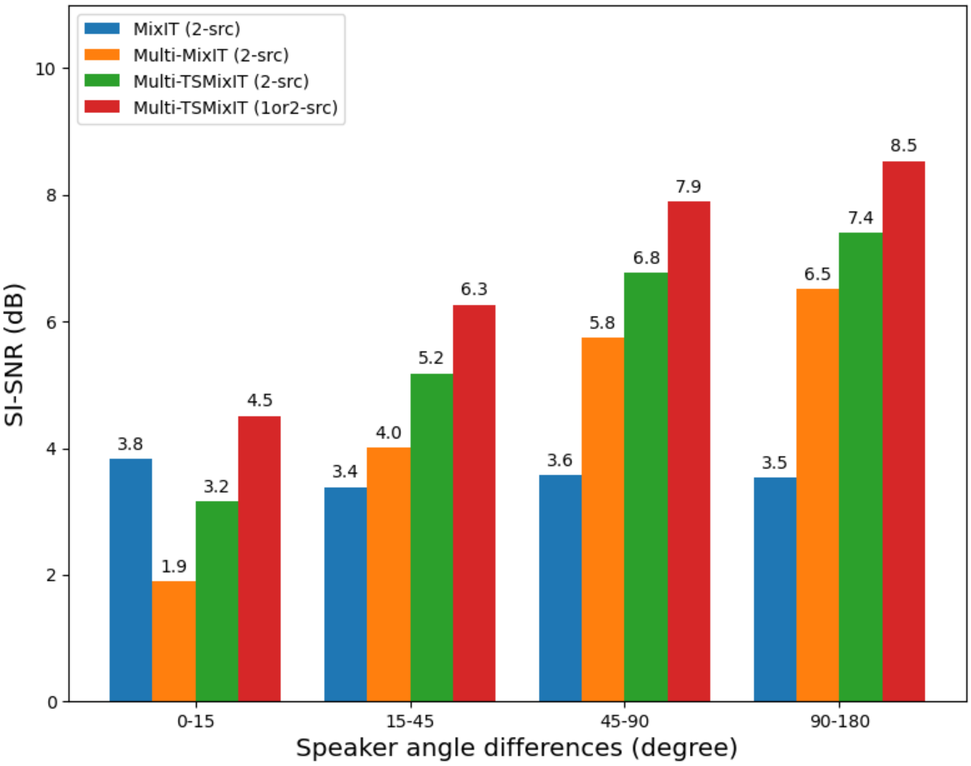


Figure 6.4: Performance of baseline and proposed multi-channel unsupervised separation systems versus angular separation between the two speakers.

audio sample, which is either the spectral cue or the spatial cue. Thirdly, the results show that the TS-MixIT method improves the performance over the MixIT method in all conditions. Further improvements in all conditions are obtained by accessing single-speaker signal as training targets, denoted as Multi-TSMixIT (1or2-src).

Then, we assess the ASR performance by using the unsupervised separation model as front-end in the noisy condition. The used acoustic model is trained using clean WSJ0/WSJ1 SI-284 data plus WHAMR! noisy and reverberant single-speaker speech. Both speech signals separated from a separation model are decoded by the acoustic model to compute the WER.

Table 6.4 presents the ASR performance. The top part in the table shows the performance of separation models trained with 2-source mixtures. The MixIT model is able to reduce the WER from 79.1% with the unprocessed mixture to 47.5%. Although the TS-MixIT improves the SI-SNR score, it only yields a slight gain to the ASR performance compared to the vanilla MixIT. This is expected since the overall separation performance is still poor, i.e., around 6 dB SI-SNR improvement. The mixtures are still not well separated and many distortions exist in

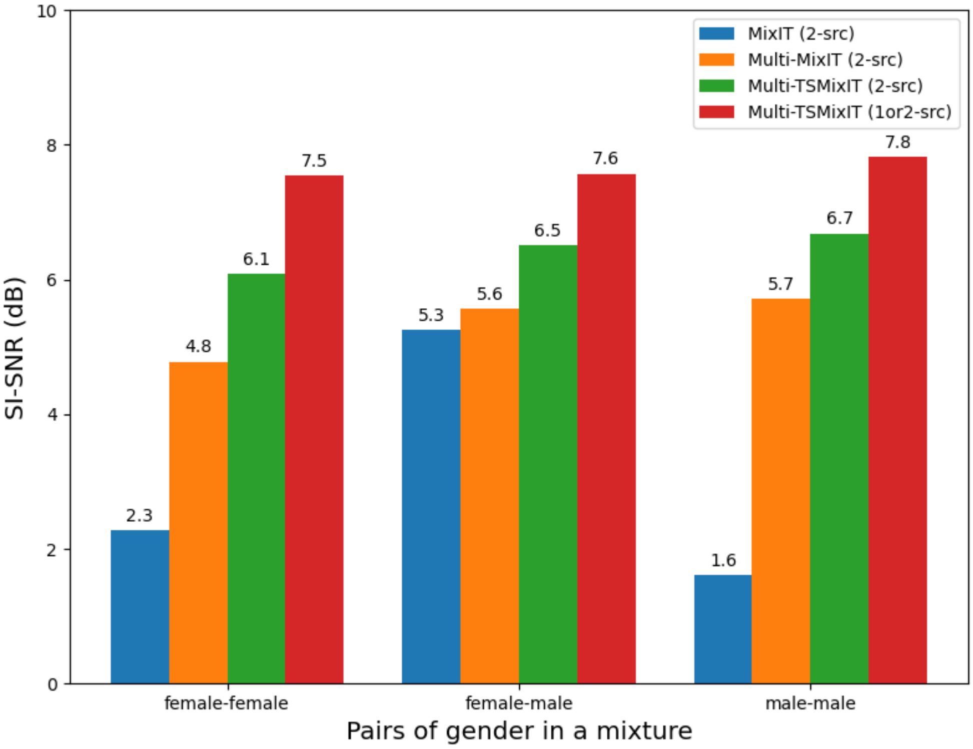


Figure 6.5: Performance of multi-channel unsupervised separation under three pairs of gender conditions.

the separated signals, which will reduce the ASR performance. The middle part presents the results by using separation systems trained with 1or2-source mixtures. Accessing single-speaker speech improves the quality of reconstructed signals and hence further reduces the WER. In addition, the TS-MixIT outperforms the MixIT with 1or2-source mixtures, reducing the WER from 44.6% to 42.1%. It suggests that the TS-MixIT approach could provide a significant benefit when the teacher MixIT model can provide separated samples with good quality. In the bottom part, the recognition performance of the fully supervised separation model is presented for comparison. It can be noticed that the WER achieved by the TS-MixIT approach with 1or2-source mixtures is only 3% higher than that achieved by the fully supervised approach, which is 39.3%. This demonstrates that the proposed unsupervised learning approach can efficiently train a separation model to be used an ASR front-end in noisy conditions.

Table 6.4: Speech recognition results using unsupervised multi-channel separation models with WHAMR! data. The separation model is applied to noisy reverberant two-speaker mixtures.

System	**#nchs**	**WER(%)**
Mixture	-	79.1
MixIT (2-src)	2	47.5
TS-MixIT (2-src)	2	47.2
MixIT (1or2-src)	2	44.6
TS-MixIT (1or2-src)	2	42.1
Multi-TCN	2	39.3
Noisy single-speech	-	19.3
Reverberant single-speech	-	11.4

6.5.2 Semi-supervised speech separation

Anechoic condition

The effect of model fine-tuning and knowledge distillation is assessed next. We first use WSJ0-2mix to conduct experiments in the anechoic condition. The TS-MixIT model trained with 1or2-source is fine-tuned with the selected 10% of the official training data. For the knowledge distillation experiment, the fine-tuned model processes the mixtures in the unsupervised training data to provide pseudo-targets to train a new student model, i.e., a DPRNN model [Luo et al., 2020].

Results in Table 6.5 show that the fine-tuned model achieves 13.2 dB of SI-SNR gain (row 3) and outperforms both the student model (row 1) and the supervised model trained with the same amount of data (row 6). This indicates that unsupervised pre-training is essential to improve the model generalisation ability. Through knowledge distillation, it can exploit the knowledge from the fine-tuned teacher model more effectively and achieves a gain of 14.3 dB SI-SNR (row 4). It is also observed that directly training the DPRNN with MixIT using the energy selection criterion leads to poor performance (line 5 in Table 6.5), indicating that the teacher-student approach is important for successfully exploiting other model architectures.

Next, we assess the semi-supervised separation models when applied as front-end for the overlapping speech recognition. Table 6.6 presents the ASR performance in the anechoic condition without noise. By fine-tuning the TS-MixIT model (1or2-src) with a small amount of su-

Table 6.5: Results with model fine-tuning and knowledge distillation for anechoic single-channel separation

Method	**Architecture**	**SI-SNRi (dB)**
TS-MixIT (Proposed)	Conv-TasNet	12.6
+Distill	DPRNN	12.9
+Fine_tuning (10%)	Conv-TasNet	13.2
+Fine_tuning (10%)+Distill	DPRNN	14.3
MixIT (Energy)	DPRNN	4.0
Supervised (10%)	Conv-TasNet	11.3
Supervised (10%)	DPRNN	11.7
Supervised (100%)	Conv-TasNet	15.3
Supervised (100%)	DPRNN	18.8

Table 6.6: Speech recognition performance of semi-supervised separation under the anechoic condition without noise.

System	**WER(%)**
Mixture	79.6
MixIT (1or2-src)	26.7
TS-MixIT (1or2-src)	26.2
+Fine_tuning (10%)	23.7
+Fine_tuning (10%)+Distill	22.5
Supervised (100%)	19.3
Oracle	9.1

pervised data, the WER is reduced from 26.2% to 23.7%. The separation model with knowledge distillation achieves 22.5% WER, compared to 23.7% for the fine-tuned model. These results indicate that the improvements in the separation model that are obtained by fine-tuning and knowledge distillation can benefit the overlapping speech recognition in the anechoic condition.

Noisy and reverberant condition

Here, we use the WHAMR! data to evaluate how the semi-supervised approaches perform in the noisy and reverberant conditions. For the fine-tuning, the noisy and reverberant mixtures are set as the input to the separation model and the reverberant sources are used as targets. In this case, the fine-tuning forces the separation model to learn both denoising and speaker separation.

Table 6.7: Results with model fine-tuning and knowledge distillation for noisy multi-channel separation

Method	**Architecture**	**SI-SNRi (dB)**
TS-MixIT (Proposed)	Conv-TasNet	7.7
+Fine_tuning (10%)	Conv-TasNet	9.2
+Fine_tuning (10%)+Distill	U-ConvBlock	9.7
Supervised (10%)	Conv-TasNet	8.5
Supervised (10%)	U-ConvBlock	8.5
Supervised (100%)	Conv-TasNet	11.1
Supervised (100%)	U-ConvBlock	12.1

Then, the fine-tuned model processes mixtures in the training set to generate separated sources, which are used to train a U-ConvBlock based multi-channel separation model.

Table 6.7 reports the SI-SNR results achieved by multi-channel separation models. With 10% supervised data for fine-tuning (line 2 in Table 6.7), the model achieves a large gain over the unsupervised student model. This can be explained as the unsupervised pre-trained model is encouraged to be more specific in the denoising and separation tasks with the supervised data. In addition, the fine-tuned model outperforms the supervised model trained with the same amount of paired data (line 4 in Table 6.7), indicating that the model can benefit from the unsupervised pre-training. Next, the knowledge distillation method is assessed. With 100% supervised data, the U-ConvBlock based multi-channel system achieves 12.1 dB, compared to 11.1 dB achieved by the TCN based model, indicating that the U-ConvBlock system is a better design for the separation task. It is observed that the knowledge distillation yields an additional 0.5 dB SI-SNR improvement compared to the fine-tuned model. This demonstrates that the knowledge distillation can exploit different network architectures in noisy environments.

To understand how spatial diversity variation of overlapped speakers impacts on the semi-supervised separation system, the separation performance is measured in conditions with various speaker angular separation. The test set is split into four subsets, each of which contains a unique range of angular separation. Figure 6.6 presents the SI-SNR improvement achieved by four separation systems in each subset. We first observe that the fine-tuning improve the performances over the unsupervised TS-MixIT model in all angular difference conditions, however, it provides a larger gain in the small angular difference condition (0-15 degree) than those

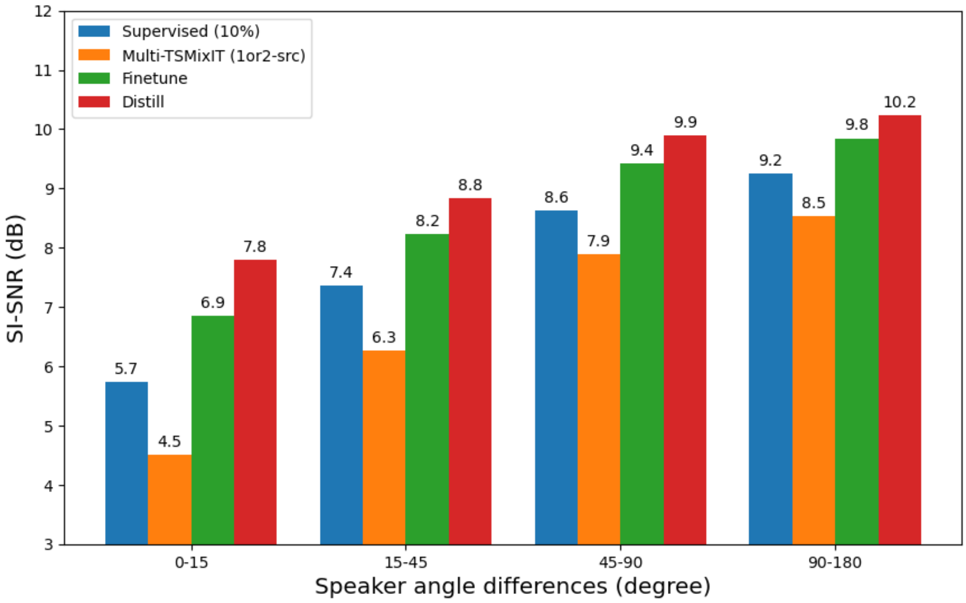


Figure 6.6: Performance of baseline and proposed multi-channel semi-supervised separation systems versus angular difference between the two speakers.

achieved in large angular difference condition (larger than 45 degree). This suggests that the supervised data could be helpful specifically in situations where an unsupervised model fails. Secondly, the knowledge distillation approach provides consistent gains over the fine-tuned model in all angular separation conditions.

Here, we investigate how speaker similarities impact the semi-supervised separation systems in the noisy condition. Figure 6.7 shows that the fine-tuning approach exploits the supervised data to provide significant improvements in all mixture gender conditions. Furthermore, by exploiting an advanced separation network architecture, the knowledge distillation approach yields consistent gains over the fine-tuned model in all mixture gender conditions.

Next, the effect of the semi-supervised separation methods on the speech recognition performance is assessed. The ASR decoding uses the acoustic model trained with clean WSJ0/WSJ1 SI-284 data plus WHAMR! noisy and reverberant single-speaker speech. The results are presented in Table 6.8. The fine-tuning and knowledge distillation approaches achieve 47.6% and 44.7% of WER, respectively. These results are worse than that achieved by the unsupervised TS-MixIT model (1or2-src), and is in contrast to the trend of SI-SNR measurement. This is possibly due to the lack of supervised data for model fine-tuning, which changes the task from separation to joint denoising and separation. It is assumed that, when the amount of supervised

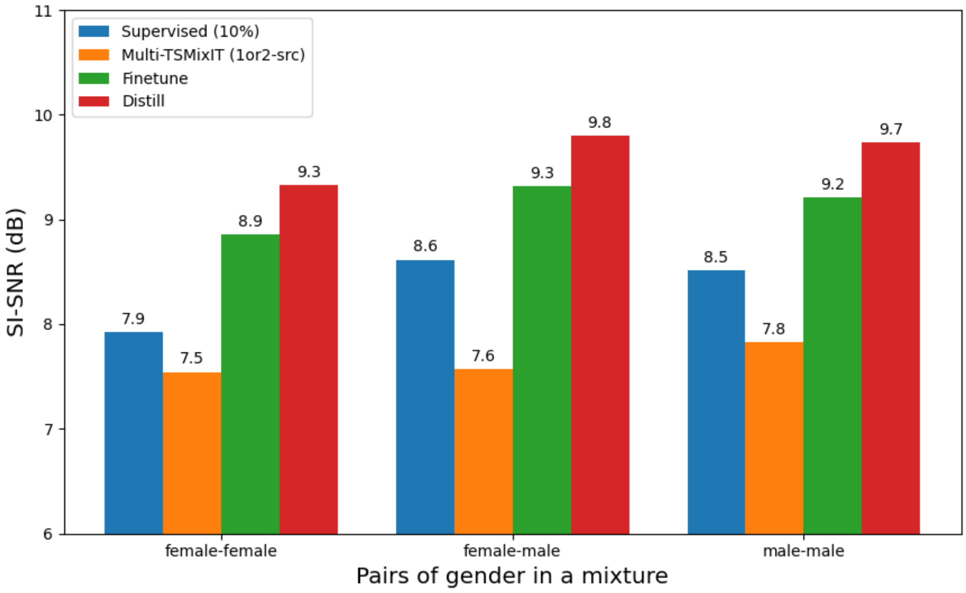


Figure 6.7: Performance of multi-channel semi-supervised separation under three pairs of gender conditions.

data is relatively small, the separation model may introduce more distortions to the separated signals than enhancement. Since the distortion is unseen during training the AM1, this may cause a severe mismatch problem that hurts the ASR performance.

Lastly, we investigate an approach to reduce the impact of distortions on the ASR performance. The distortion introduced by the separation process causes mismatched conditions between training and test data for the acoustic model. To remove the mismatched condition, a multi-condition training method is used to train acoustic models with additional enhanced data. The enhanced data is obtained by applying a separation model to the noisy and reverberant mixtures in the WHAMR! training set. This will improve the robustness of a DNN based acoustic model to diverse distortions in the test data [Kinoshita et al., 2016]. Three additional acoustic models are trained, each of which will match a separation model, namely TS-MixIT (1or2-src), fine-tuned model, and the knowledge distillation model. The matched acoustic models are labelled as AM2.

Table 6.9 compares the performances achieved by the mismatched ASR model (AM1) and the ASR model under multi-conditioning training (AM2). For all separation models, a large accuracy improvement is achieved by using the matched acoustic model. Interestingly, we observe that, using the AM2, the distillation system achieves lower WER compared with the

Table 6.8: Speech recognition results with noisy and reverberant two-speaker mixtures.

System	**#nchs**	**WER(%)**
Mixture	-	79.1
MixIT (1or2-src)	2	44.6
TS-MixIT (1or2-src)	2	42.1
+Fine_tuning (10%)	2	47.6
+Fine_tuning (10%)+Distill	2	44.7
Multi-TCN (100%)	2	39.3
Multi-UConv (100%)	2	34.9
Noisy single-speech	-	19.3
Reverberant single-speech	-	11.4

Table 6.9: Speech recognition results with mismatched and matched acoustic models. AM1 and AM2 denote the mismatched and matched acoustic models, respectively.

System	**#nchs**	**WER(%)**	
		AM1	AM2
TS-MixIT (1or2-src)	2	42.1	32.9
+Fine_tuning (10%)	2	47.6	30.9
+Fine_tuning (10%)+Distill	2	44.7	28.1
Noisy single-speech	-	19.3	19.3

unsupervised TS-MixIT model (1or2-src). This indicates that the fine-tuned model injects more artificial distortions to the separated signals than the unsupervised separation model. Although multi-condition training for the acoustic model can reduce the impact of distortions, further investigation is needed on how to control the unexpected distortions introduced from time-domain separation models.

6.6 Summary

This chapter has presented novel unsupervised and semi-supervised approaches for end-to-end speech separation. The first piece of work is to use a teacher-student learning framework to improve an existing unsupervised approach, i.e. mixture invariant training criterion (MixIT). It has been shown that the MixIT model suffers an over-separation problem caused by mismatches between training and testing conditions. By using the MixIT model as a teacher to train a student model with mixture signals that are matched to the testing condition, the proposed method resolves the teacher's over-separation problem.

The proposed unsupervised learning framework can be applied to both single- and multi-channel speech separation systems. Notable, the proposed approach significantly improves the multi-channel separation performance over the MixIT framework in realistic conditions including both reverberation and additional non-speech noise sources. The separation model under the proposed unsupervised training framework can effectively serve as a front-end for speech recognition with overlapping speech in noisy and reverberant conditions. Based on simulated data, the proposed unsupervised method reduced the WER from 79.1% of unprocessed mixture to 42.1%. By conducting multi-condition training to the acoustic model to improve its robustness to the distortions introduced by the separation process, the speech recognition performance is further improved to 32.9%.

The second piece of work is a novel semi-supervised approach that considers the situation where a limited amount of supervised data is available. The proposed semi-supervised framework contains three stages: 1) unsupervised pre-training, 2) supervised fine-tuning, and 3) model distillation. It has been shown that an unsupervised pre-trained separation model can be improved via fine-tuning using limited supervised data. Then, the fine-tuned model can be further improved via model distillation, which effectively exploits different separation network architectures for unsupervised pre-training and the later semi-supervised learning. The separation performance of the semi-supervised method is comparable to that of a fully supervised system using ten times the amount of supervised data. By exploiting limited supervised data, the semi-supervised approach further improves the overlapping speech recognition performance, achieving 28.1% of WER, compared to 32.9% achieved by the unsupervised approach.

Chapter 7

Conclusion

This thesis has focused on the challenge of distant speech recognition in multi-talker scenarios, in which speaker overlaps significantly degrade the speech recognition performance. To improve overlapping speech recognition performance, efforts have been made to explore speech separation technologies. Motivated by recent progress achieved by deep-learning based single-channel time-domain separation approaches, this work has advanced the separation technique by exploring the use of multiple recordings, speaker identity information, and unsupervised learning. The methods that have been developed have been shown to improve speech separation systems and benefit distant speech recognition in noisy multi-talker scenarios.

The research has been organised around a sequence of four specific research questions.

- Research question 1. The extent of overlap: What is the extent of overlapping speech in casual conversations and what impacts overlapping speech can have on current state-of-the-art ASR systems?
- Research question 2. Processing multi-channel signals: How best to effectively exploit multi-channel information in an end-to-end time-domain speech separation system?
- Research question 3. Using speaker identity: How best to exploit speaker identity information in an end-to-end time-domain multi-channel extraction system?
- Research question 4. Adapting to real data: How to build a speech separation network in situations where the amount of supervised training data is limited? This question has tried to bridge the gap between simulated and real scenarios.

The final chapter, will conclude by first reviewing the main thesis contributions with respect to the above four questions. Section 7.2 will then provide suggestions for potential future research directions.

7.1 Contributions

Analysis of conversational speech recordings

The first contribution is a thorough analysis about the extent of overlapping speech in a natural multi-talker party scenario. This analysis has been performed to address the first research question, i.e., to establish the significance of the overlapping speech problem. Analysis conducted on the CHiME-5 data has shown that speaker overlaps can account for as much as 30% of time period of conversational speech and provides evidence that the overlapping speech has a significant negative impact on ASR performance. Experiments presented in Section 3.4 have shown that the overlapping speech is a key factor for the high WER of 74.14% achieved in distant speech recognition on the CHiME-5 data. It has been found that when the proportion of overlapped speech in an utterance increases, the ASR accuracy drops.

As initial attempts, two approaches, device selection and speech separation, have been applied to multi-talker scenarios to address the problem caused by overlapping speech. It was found that, even in the CHiME-5 situation in which up to six devices are distributed within an often open-plan living space, the device selection approach provided limited gains to distant speech recognition in multi-talker real home environments. Experiments in Section 3.5 have shown that selecting the best device based on a speech intelligibility measurement achieved absolute 4% WER reduction on the CHiME-5 data. Even if the device is selected based purely on recognition performance, (i.e., using knowledge of the transcript to provide an upper-bound for the benefit of the approach), then the WER reduction was only absolute 14% (from 74% to 60%). This finding suggests that the problem caused by overlapping speech cannot be fully addressed by adding more devices to the living space and selecting one that has a good speech signal quality. Aspects of this work were published in Xiong et al. [2018].

However, in contrast, even an existing baseline speech separation method provided a substantial improvement compared with the unprocessed signal and greatly outperformed the channel selection method. Evaluation presented in Section 3.6 has shown that an existing speech separation method yielded absolute 17% WER reduction compared to the channel selection approach. This suggests that speech separation is the key solution to overlapping speech recognition.

Blind time-domain multi-channel separation

Motivated by the observation that speech separation effectively improved multi-talker speech recognition in noisy environments, the thesis has focused on the development of more powerful speech separation techniques. The key strategy has been to focus on exploiting multi-channel recordings and recent advances in deep learning techniques, i.e. the end-to-end time-domain framework for speech separation. The second research question asks how can multi-channel information be effectively exploited in an end-to-end time-domain speech separation system.

A fully-convolutional neural network structure for time-domain multi-channel speech separation has been successfully designed in Chapter 4. It is argued that using conventional time-frequency domain spatial features is not an optimal choice for the time-domain speech separation system. Therefore, this work has proposed to use neural networks to extract spatial features from time-domain signals. Specifically, a 2-dimensional convolutional layer has been demonstrated to be able to efficiently learn spatial features from multi-recordings in the time-domain. In Section 4.5.1, experiments using simulated reverberant multi-channel recordings (spatialised WSJ0-2mix) have shown that the learned spatial features are suitable for the time-domain separation system and lead to an improvement of separation and recognition performance over conventional time-frequency domain spatial features.

The effect of reverberation on the feature extraction within the end-to-end approach has been investigated. It was found that the neural network based spectral and spatial encoders are sensitive to the reverberation and the quality of extracted features are degraded in reverberant environments. This influence damages the quality of the separated signals and as a result degrades the ASR performance. Applying dereverberation methods to the noisy mixture as a preprocessing step before the separation reduces the effect caused by reverberation on the proposed end-to-end time-domain multi-channel system. Experiments in Section 4.5.2 have shown that this dereverberation stage improves the quality of the separated speech signals and benefits a subsequent speech recognition system.

The design of network architectures has been explored to better exploit multi-channel information in a separation system. Two architecture designs, dilated convolution and subsampling for convolutional networks, have been investigated. In Section 4.5.3, experiments using simulated two-channel recordings in noisy and reverberant environments (WHAMR! data) have shown that subsampling operations are more effective at increasing the temporal modelling capacity of a time-domain multi-channel separation network. The improved architecture has led to improvements of both separation and recognition performances.

The deep-learning based separation system introduces many artificial distortions to the separated signals. To reduce the distortions, a two-stage separation procedure has been proposed, which re-used the time-domain multi-channel network as an enhancement system for post-processing. In Section 4.5.4, experiments on WHAMR! data have shown that the enhanced signals achieve better speech separation and recognition performances compared with separated signals from the single-stage separation.

Aspects of the work presented in Chapter 4 have been published in Zhang et al. [2020a].

Time-domain multi-channel speaker extraction

A speech separation system separates as many sources in a mixture as possible, however, we are often only interested in speech from a specific speaker within the mixture. For example, a system may be attending to a familiar conversational partner while not needing to process speech from interfering background speakers. With this observation in mind, the third research question asks how to exploit speaker identity information to target a separation system on a specific speaker or set of speakers.

A time-domain multi-channel extraction system has been designed, which exploits speaker identity information in terms of voice characteristics to target the extraction system on specific speakers. To efficiently fuse features from multiple modalities in an extraction system, a multi-stage speaker conditioning mechanism has been developed. Specifically, an additional speaker branch has been introduced to the extraction system to receive external speaker identity features and to fuse the identity features with the spectral and spatial features. The speaker branch is encouraged to learn a fusion strategy and to generate a new speaker identity representation that is more suitable for the extraction task than the raw speaker identity features. In Section 5.5.1, evaluated with WHAMR! data, the developed extraction system was able to exploit speaker identity features to simultaneously extract multiple speakers, and improved overlapping speech recognition compared to the baseline separation system.

The effect of the quality of speaker identity features on the speech extraction performance has been investigated. Most existing approaches use only one clean utterance from a target speaker to extract speaker identity features, which may lead to poor quality. To improve the quality of the speaker identity representation, this work has proposed to average speaker features extracted from multiple utterances belonging to a target speaker to form global speaker identity features. In Section 5.5.2, evaluation on WHAMR! data has show that the quality of the speaker identity features can affect the speaker extraction performance. The proposed global features improved the extraction performance over features extracted from a single utterance and yielded better speech recognition performance.

A further study has compared single-speaker extraction and simultaneous multi-speaker extraction. It was observed in Section 5.5.3 that extracting one target speaker each time yielded better separation and recognition performances. It is suspected that the multi-speaker extraction is a more challenging task because the extraction system not only needs to separate speech elements from a mixture signal, but also associate these elements to different speakers, which may lead to identity switches or lost tracks. While in a single-speaker extraction system, there is no speaker association problem. Reasons for the difference in the performance of these two strategies are not fully understood and require further investigation.

Early outputs of this part of the work have been published in Zhang et al. [2021b].

Unsupervised and semi-supervised end-to-end separation

Speech separation systems trained on simulated data have a mismatch problem when applied to real environments. In real scenarios, large amounts of supervised training data is hard to collect, which makes it difficult to conduct supervised training for a speech separation network. The fourth research question asks how to build deep-learning based separation systems in situations where large amounts of unsupervised training data is provided and only a small amount of supervised training data is available.

As discussed in Section 2.5, existing unsupervised and semi-supervised learning approaches have been designed for developing end-to-end speech separation systems with only unlabelled mixture signals. However, they still suffer a mismatch problem between training and testing conditions. To address this problem, a novel unsupervised approach has been developed. The proposed method combines teacher-student learning [Lam et al., 2020] and an existing unsupervised learning approach, i.e., the mixture invariant training (MixIT) framework [Wisdom et al., 2020]. The MixIT model is trained with an artificial mixture of mixtures, so this artificial mixture contains more sources than a typical mixture evaluated at test time. This training mismatch results in an over-separation problem in the MixIT model. The proposed method employed a teacher-student learning approach that uses the MixIT model as a teacher to generate pseudo-targets from original mixtures to train a student model with a smaller number of output channels. Experiments in Section 6.5.1 evaluated the proposed method with WHAMR! data and showed that, since the student model is trained directly using original mixture signals, the proposed method effectively addressed the over-separation problem caused by mismatches between the training and testing. The proposed unsupervised learning approach is able to train a multi-channel separation system to effectively improve speech recognition performance in multi-talker noisy environments.

Then, this unsupervised approach was further extended to a semi-supervised learning approach that can exploit both large amounts of unlabelled data and limited supervised training data. The proposed framework contains three stages: 1) unsupervised pretraining, 2) supervised fine-tuning, and 3) model distillation. It has been found in Section 6.5.2 that an unsupervised pre-trained separation model can be improved via fine-tuning using limited supervised data. Furthermore, it was argued that different separation network architectures should be employed for unsupervised pre-training and the later semi-supervised training. Therefore, the proposed method distilled the knowledge of the fine-tuned model to a student model which has a more task-specific network structure for speech separation. The semi-supervised separation model further improved the recognition performance over the unsupervised separation model.

Aspects of the work presented in Chapter 6 have been published in Zhang et al. [2021a].

7.2 Future research

Although the thesis has demonstrated many promising approaches that have achieved new state-of-the-art performances on the datasets that they have been evaluated on, there still remains much to be done in order to bridge gaps to real data encountered in real distant microphone applications. This section presents some potential future research directions aiming for either addressing emerging issues of the work in this study or broader applications.

Distributed microphones or microphone arrays

The multi-channel network in this work has been designed for a single microphone array. In a modern home or office, there are usually more than one microphone or array located in different places to cover the entire area, which is known as distributed microphones. This large spatial diversity provides another strong spatial cues to separate sources. Future work could explore time-domain separation methods for distributed microphones and arrays to improve the speech separation and recognition performance.

Multi-speaker extraction

As observed in Chapter 5, the performance of a multi-speaker extraction system still lags behind that of a single-speaker extraction system. When a neural network is used to identify and track multiple objects from a video recording, object association is usually a key issue that may cause identity switches and lost tracks [Keuper et al., 2020; Zheng et al., 2021]. These problems could potentially happen in a multi-speaker extraction system as well. Therefore, further analysis is needed to identify the issue that causes the performance gap between single- and multi-speaker extraction systems. This analysis will guide new system designs to stabilise the multi-speaker extraction process.

More unsupervised training data

The proposed unsupervised approach in this work is evaluated with a medium size of unlabelled data and has not fully exploited large amounts of noisy signals available in the real world. This seems to be on the contrary to the original motivation of unsupervised learning. It will be interesting to see if increasing the amount of unlabelled data could improve the unsupervised model robustness and improve the subsequent semi-supervised model. Future work could exploit large amounts of unlabelled data from multiple domains to train an unsupervised separation model.

Adaptation to real scenarios

In this work, most of the developed systems are evaluated with fully-overlapped speech that is artificially generated by using plain speech, simulated room impulse responses, and pre-recorded noise. The mixture invariant training based unsupervised approach developed in Chapter 6 has successfully built a speech separation system by accessing only noisy signals, providing a potential solution to training a neural network based speech separation system with real recordings. However, this unsupervised method has only been evaluated using synthetic data. In real settings, acoustic conditions can be different from that of the simulated settings, and the effectiveness of the unsupervised method for real data remains to be verified.

At the time of writing, a recent work has shown that the mixture invariant training can be extended to exploit both unpaired clean speech and real noisy speech to effectively improve recognition performance of single-talker single-channel noisy speech [Zhang et al., 2022]. Motivated by this, future work could investigate how mixture invariant training can be explored to exploit both unpaired clean speech and real noisy speech in real multi-talker multi-channel scenarios.

Bibliography

Allen, J. B. and Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950.

Alumäe, T. and Kurimo, M. (2010). Efficient estimation of maximum entropy language models with n-gram features: an srilm extension. In *Proc. Interspeech 2010*.

Anguera, X., Wooters, C., and Hernando, J. (2007). Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2011–2022.

Araki, S., Makino, S., Blin, A., Mukai, R., and Sawada, H. (2004). Underdetermined blind separation for speech in real environments with sparseness and ICA. In *2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Araki, S., Okada, M., Higuchi, T., Ogawa, A., and Nakatani, T. (2016). Spatial correlation model based observation vector clustering and mvdr beamforming for meeting recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Bai, S., Kolter, J. Z., and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *ArXiv*, abs/1803.01271.

Bando, Y., Sasaki, Y., and Yoshii, K. (2019). Deep bayesian unsupervised source separation based on a complex gaussian mixture model. In *2019 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*.

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

Barker, J., Marxer, R., Vincent, E., and Watanabe, S. (2017). The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes. *Computer Speech & Language*, 46:605–626.

Barker, J., Vincent, E., Ma, N., Christensen, H., and Green, P. D. (2013). The PASCAL CHiME speech separation and recognition challenge. *Computer Speech & Language*, 27:621–633.

Barker, J., Watanabe, S., Vincent, E., and Trmal, J. (2018). The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines. In *Proc. Interspeech 2018*.

Benesty, J., Chen, J., Huang, Y., and Dmochowski, J. P. (2007). On Microphone-Array Beamforming From a MIMO Acoustic Signal Processing Perspective. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1053–1065.

Boeddecker, C., Heitkaemper, J., Schmalenstroeer, J., Drude, L., Heymann, J., and Haeb-Umbach, R. (2018). Front-end processing for the chime-5 dinner party scenario. In *Proc. 5th International Workshop on Speech Processing in Everyday Environments.*

Brungart, D. S. (2005). Informational and energetic masking effects in multitalker speech perception. In *Speech separation by humans and machines*, pages 261–267. Springer US.

Cauchi, B., Kodrasi, I., Rehr, R., Gerlach, S., Jukic, A., Gerkmann, T., Doclo, S., and Goetze, S. (2015). Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech. *EURASIP Journal on Advances in Signal Processing*, 2015:1–12.

Cetin, O. and Shriberg, E. (2006). Speaker overlaps and ASR errors in meetings: Effects before, during, and after the overlap. In *2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Chang, X., Qian, Y., Yu, K., and Watanabe, S. (2019). End-to-end monaural multi-speaker ASR system without pretraining. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Chen, S., Wu, Y., Chen, Z., Li, J., Wang, C., Liu, S., and Zhou, M. (2021). Continuous speech separation with conformer. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. (2020a). Big self-supervised models are strong semi-supervised learners. In *Proc. NeurIPS 2020*.

Chen, Z., Droppo, J., Li, J., Xiong, W., Chen, Z., Droppo, J., Li, J., and Xiong, W. (2018a). Progressive joint modeling in unsupervised single-channel overlapped speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 26(1):184–196.

Chen, Z., Luo, Y., and Mesgarani, N. (2017). Deep attractor network for single-microphone speaker separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Chen, Z., Xiao, X., Yoshioka, T., Erdogan, H., Li, J., and Gong, Y. (2018b). Multi-channel overlapped speech recognition with location guided speech extraction network. *2018 IEEE Spoken Language Technology Workshop (SLT)*.

Chen, Z., Yoshioka, T., Lu, L., Zhou, T., Meng, Z., Luo, Y., Wu, J., and Li, J. (2020b). Continuous speech separation: Dataset and analysis. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cooke, M., Hershey, J. R., and Rennie, S. J. (2010). Monaural speech separation and recognition challenge. *Computer Speech & Language*, 24:1–15.

Cosentino, J., Pariente, M., Cornell, S., Deleforge, A., and Vincent, E. (2020). Librimix: An open-source dataset for generalizable speech separation. *arXiv: Audio and Speech Processing*.

Cox, H., Zeskind, R. M., and Owen, M. M. (1987). Robust adaptive beamforming. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(10):1365–1376.

D'efossez, A., Usunier, N., Bottou, L., and Bach, F. (2019). Music source separation in the waveform domain. *ArXiv*, abs/1911.13254.

Dehak, N., Kenny, P., 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.

Delcroix, M., Ochiai, T., Zmolikova, K., Kinoshita, K., Tawara, N., Nakatani, T., and Araki, S. (2020). Improving speaker discrimination of target speech extraction with time-domain Speakerbeam. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Delcroix, M., Žmolíková, K., Kinoshita, K., Ogawa, A., and Nakatani, T. (2018). Single channel target speaker extraction and recognition with speaker beam. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Delcroix, M., Žmolíková, K., Ochiai, T., Kinoshita, K., Araki, S., and Nakatani, T. (2019). Compact network for speakerbeam target speaker extraction. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Delfarah, M. and Wang, D. (2018). Recurrent neural networks for cochannel speech separation in reverberant environments. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Delfarah, M. and Wang, D. (2019). Deep learning for talker-dependent reverberant speaker separation: An empirical study. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 27(11):1839–1848.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Drude, L., Hasenklever, D., and Haeb-Umbach, R. (2019a). Unsupervised training of a deep clustering model for multichannel blind source separation. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Drude, L., Heitkaemper, J., Boeddeker, C., and Haeb-Umbach, R. (2019b). SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition. *ArXiv*, abs/1910.13934.

Drude, L., Heymann, J., and Haeb-Umbach, R. (2019c). Unsupervised Training of Neural Mask-Based Beamforming. In *Proc. Interspeech 2019.*

Du, J., Tu, Y., Xu, Y., Dai, L., and Lee, C.-H. (2014). Speech separation of a target speaker based on deep neural networks. In *2014 12th International Conference on Signal Processing (ICSP).*

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W., and Rubinstein, M. (2018). Looking to listen at the cocktail party. *ACM Transactions on Graphics (TOG)*, 37:1 – 11.

Erdogan, H., Hershey, J., Watanabe, S., Mandel, M. I., and Roux, J. L. (2016). Improved mvdr beamforming using single-channel mask prediction networks. In *Proc. Interspeech 2016*.

Erdogan, H., Hershey, J. R., Watanabe, S., and Roux, J. L. (2015). Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Erhan, D., Courville, A. C., Bengio, Y., and Vincent, P. (2010). Why does unsupervised pre-training help deep learning? *The Journal of Machine Learning Research*, 11:625–660.

Fox, C., Liu, Y., Zwyssig, E., and Hain, T. (2013). The sheffield wargames corpus. In *Proc. Interspeech 2013*.

Ge, M., Xu, C., Wang, L., Siong, C. E., Dang, J., and Li, H. (2020). Spex+: A complete time domain speaker extraction network. In *Proc. Interspeech 2020*.

Griffiths, L. J. and Jim, C. W. (1982). An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30:27–34.

Gu, R., Wu, J., Zhang, S.-X., Chen, L., Xu, Y., Yu, M., Su, D., Zou, Y., and Yu, D. (2019). End-to-end multi-channel speech separation. *arXiv preprint arXiv:1905.06286*.

Guerrero, C., Tryfou, G., and Omologo, M. (2018). Cepstral distance based channel selection for distant speech recognition. *Computer Speech & Language*, 47:314–332.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech 2020*.

Han, K., Wang, Y., Wang, D., Woods, W. S., Merks, I., and Zhang, T. (2015). Learning spectral mapping for speech dereverberation and denoising. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(6):982–992.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Heitkaemper, J., Jakobeit, D., Boeddeker, C., Drude, L., and Haeb-Umbach, R. (2020). Demystifying tasnet: A dissecting approach. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Hershey, J. R., Rennie, S. J., Olsen, P. A., and Kristjansson, T. T. (2010). Super-human multi-talker speech recognition: A graphical modeling approach. *Computer Speech & Language*, 24:45–66.

Heymann, J., Drude, L., and Haeb-Umbach, R. (2016). Neural network based spectral mask estimation for acoustic beamforming. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. (2014). Deep learning for monaural speech separation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Isik, Y., Le Roux, J., Chen, Z., Watanabe, S., and Hershey, J. R. (2016). Single-channel multi-speaker separation using deep clustering. In *Proc. Interspeech 2016*.

Ito, N., Araki, S., and Nakatani, T. (2016). Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing. In *2016 24th European Signal Processing Conference (EUSIPCO)*.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al. (2003). The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ji, X., Yu, M., Zhang, C., Su, D., Yu, T., Liu, X., and Yu, D. (2020). Speaker-aware target speaker enhancement by jointly learning with speaker embedding extraction. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

John W. Du Bois, Wallace L. Chafe, C. M. and Thompson, S. A. (2000). Santa barbara corpus of spoken american english, parts 1-4. *Philadelphia: Linguistic Data Consortium*.

Kanda, N., Boeddeker, C., Heitkaemper, J., Fujita, Y., Horiguchi, S., Nagamatsu, K., and Haeb-Umbach, R. (2019). Guided Source Separation Meets a Strong ASR Backend: Hitachi/Paderborn University Joint Investigation for Dinner Party ASR. In *Proc. Interspeech 2019*.

Kellermann, W. (2008). *Beamforming for Speech and Audio Signals*, pages 691–702. Springer New York, New York, NY.

Keuper, M., Tang, S., Andres, B., Brox, T., and Schiele, B. (2020). Motion segmentation & multiple object tracking by correlation co-clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):140–153.

Khabbazibasmenj, A., Vorobyov, S. A., and Hassanien, A. (2012). Robust adaptive beamforming based on steering vector estimation with as little as possible prior information. *IEEE Transactions on Signal Processing*, 60(6):2974–2987.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kinoshita, K., Delcroix, M., Gannot, S., Habets, E. A. P., Häb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., Sehr, A., and Yoshioka, T. (2016). A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, 2016:1–19.

Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., and Maas, R. (2013). The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proc. Interspeech 2015*.

Kolbæk, M., Tan, Z.-H., Jensen, S. H., and Jensen, J. (2020). On loss functions for supervised monaural time-domain speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:825–838.

Kolbæk, M., Yu, D., Tan, Z.-H., and Jensen, J. (2017). Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1901–1913.

Lam, M. W., Wang, J., Su, D., and Yu, D. (2020). Mixup-breakdown: a consistency training method for improving generalization of speech separation models. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Lea, C., Vidal, R., Reiter, A., and Hager, G. D. (2016). Temporal convolutional networks: A unified approach to action segmentation. In *2016 European Conference on Computer Vision*.

Li, W., Zhang, P., and Yan, Y. (2019). Target Speaker Recovery and Recognition Network with Average x-Vector and Global Training. In *Proc. Interspeech 2019*.

Luo, W., Li, Y., Urtasun, R., and Zemel, R. S. (2016). Understanding the effective receptive field in deep convolutional neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*.

Luo, Y., Ceolini, E., Han, C., Liu, S.-C., and Mesgarani, N. (2019). Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Luo, Y., Chen, Z., Hershey, J. R., Roux, J. L., and Mesgarani, N. (2017). Deep clustering and conventional networks for music separation: Stronger together. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Luo, Y., Chen, Z., and Yoshioka, T. (2020). Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Luo, Y. and Mesgarani, N. (2018a). Real-time single-channel dereverberation and separation with time-domain audio separation network. In *Proc. Interspeech 2018*.

Luo, Y. and Mesgarani, N. (2018b). Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Luo, Y. and Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266.

Maciejewski, M., Sell, G., Fujita, Y., García-Perera, L. P., Watanabe, S., and Khudanpur, S. (2019). Analysis of robustness of deep single-channel speech separation using corpora constructed from multiple domains. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

Maciejewski, M., Shi, J., Watanabe, S., and Khudanpur, S. (2021). 2021 training noisy single-channel speech separation with noisy oracle sources: A large gap and a small step. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Maciejewski, M., Wichern, G., McQuinn, E., and Roux, J. L. (2020). WHAMR!: Noisy and reverberant single-channel speech separation. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Mandel, M. I., Weiss, R. J., and Ellis, D. P. W. (2010). Model-based expectation-maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):382–394.

Manohar, V., Chen, S.-J., Wang, Z., Fujita, Y., Watanabe, S., and Khudanpur, S. (2019). Acoustic modeling for overlapping speech recognition: Jhu chime-5 challenge system. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Medennikov, I., Korenevsky, M., Prisyach, T., Khokhlov, Y., Korenevskaya, M., Sorokin, I., Timofeeva, T., Mitrofanov, A., Andrusenko, A., Podluzhny, I., Laptev, A., and Romanenko, A. (2020). The STC System for the CHiME-6 Challenge. In *Proc. 6th International Workshop on Speech Processing in Everyday Environment*.

Menne, T., Sklyar, I., Schlüter, R., and Ney, H. (2019). Analysis of deep clustering as preprocessing for automatic speech recognition of sparsely overlapping speech. In *Proc. Interspeech 2019*.

Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S. M., Tyagi, A., Casas, J. R., Turmo, J., Cristoforetti, L., Tobia, F., et al. (2007). The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language resources and evaluation*, 41(3):389–407.

Nachmani, E., Adi, Y., and Wolf, L. (2020). Voice separation with an unknown number of multiple speakers. In *Proceedings of the 37th International Conference on Machine learning (ICML)*.

Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., and Juang, B.-H. (2010). Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1717–1731.

Neverova, N., Wolf, C., Taylor, G. W., and Nebout, F. (2016). Moddrop: Adaptive multimodal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706.

Novak, J. R., Minematsu, N., and Hirose, K. (2012). WFST-Based Grapheme-to-Phoneme Conversion: Open Source tools for Alignment, Model-Building and Decoding. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing (FSMNLP)*.

Ochiai, T., Delcroix, M., Ikeshita, R., Kinoshita, K., Nakatani, T., and Araki, S. (2020). Beam-TasNet: Time-domain Audio Separation Network Meets Frequency-domain Beamformer. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ochiai, T., Delcroix, M., Kinoshita, K., Ogawa, A., and Nakatani, T. (2019). A unified framework for neural speech separation and extraction. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Paliwal, K. K., Wójcicki, K. K., and Shannon, B. J. (2011). The importance of phase in speech enhancement. *Speech Communication*, 53:465–494.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*.

Paul, D. B. and Baker, J. M. (1992). The design for the wall street journal-based csr corpus. In *Proceedings of the Workshop on Speech and Natural Language*.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. C. (2018). FiLM: Visual Reasoning with a General Conditioning Layer. In *Proceedings of the AAAI Conference on Artificial Intelligence 2018*.

Pishdadian, F., Wichern, G., and Roux, J. L. (2020a). Finding strength in weakness: Learning to separate sounds with weak supervision. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2386–2399.

Pishdadian, F., Wichern, G., and Roux, J. L. (2020b). Learning to separate sounds from weakly labeled scenes. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohamadi, M., and Khudanpur, S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proc. Interspeech 2018*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The Kaldi speech recognition toolkit. In *2011 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Proc. Interspeech 2016*.

Ravanelli, M. and Bengio, Y. (2018). Speaker recognition from raw waveform with SincNet. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*.

Renals, S., Hain, T., and Bourlard, H. (2008). Interpretation of multiparty meetings the ami and amida projects. In *2008 Hands-Free Speech Communication and Microphone Arrays*.

Renals, S. and Swietojanski, P. (2017). Distant speech recognition experiments using the ami corpus. In *New Era for Robust Speech Recognition, Exploiting Deep Learning*.

Rickard, S. (2007). *The DUET Blind Source Separation Algorithm*, pages 217–241. Springer Netherlands.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *2015 Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L., Roomi, B., and Hall, P. (2017). English conversational telephone speech recognition by humans and machines. In *Proc. Interspeech 2017*.

Sawada, H., Araki, S., and Makino, S. (2011). Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):516–527.

Scheibler, R., Bezzam, E., and Dokmanić, I. (2018). Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*.

Seki, H., Hori, T., Watanabe, S., Le Roux, J., and Hershey, J. R. (2018). A purely end-to-end system for multi-speaker speech recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Shi, J., Xu, J., Fujita, Y., Watanabe, S., and Xu, B. (2020). Speaker-Conditional Chain Model for Speech Separation and Extraction. In *Proc. Interspeech 2020*.

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, S., Roman, N., and Wang, D. (2006). Binary and ratio time-frequency masks for robust speech recognition. *Speech Communication*, 48:1486–1501.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002)*.

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., and Zhong, J. (2021). Attention is all you need in speech separation. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Swietojanski, P., Ghoshal, A., and Renals, S. (2013). Hybrid acoustic models for distant and multichannel large vocabulary speech recognition. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136.

Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. NIPS 2017*.

Tucker, B. V. and Ernestus, M. (2016). Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. *The Mental Lexicon*, 11:375–400.

Tzinis, E., Adi, Y., Ithapu, V. K., Xu, B., and Kumar, A. (2021). Continual self-training with bootstrapped remixing for speech enhancement. *ArXiv*, abs/2110.10103.

Tzinis, E., Venkataramani, S., and Smaragdis, P. (2019). Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Tzinis, E., Venkataramani, S., Wang, Z., Subakan, C., and Smaragdis, P. (2020a). Two-step sound source separation: Training on learned latent targets. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Tzinis, E., Wang, Z., and Smaragdis, P. (2020b). Sudo rm-rf: Efficient networks for universal audio source separation. In *2020 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *ArXiv*, abs/1609.03499.

Veselý, K., Watanabe, S., Žmolíková, K., Karafiát, M., Burget, L., and Cernocký, J. (2016). Sequence summarizing neural network for speaker adaptation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Vincent, E., Barker, J., Watanabe, S., Roux, J. L., Nesta, F., and Matassoni, M. (2013). The second 'chime' speech separation and recognition challenge: Datasets, tasks and baselines. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Vincent, E., Gribonval, R., and Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:1462–1469.

Vincent, E., Watanabe, S., Nugraha, A., Barker, J., and Marxer, R. (2017). An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, 46:535–557.

von Neumann, T., Kinoshita, K., Drude, L., Boeddeker, C., Delcroix, M., Nakatani, T., and Haeb-Umbach, R. (2020). End-to-end training of time domain audio separation and recognition. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Wan, L., Wang, Q., Papir, A., and Lopez-Moreno, I. (2018). Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Wang, D. (2005). *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*, pages 181–197. Springer US.

Wang, D. and Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726.

Wang, D. and Lim, J. S. (1982). The unimportance of phase in speech enhancement. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 30(4):679–681.

Wang, P., Chen, Z., Xiao, X., Meng, Z., Yoshioka, T., Zhou, T., Lu, L., and Li, J. (2019a). Speech separation using speaker inventory. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).*

Wang, Q., Muckenhirn, H., Wilson, K., Sridhar, P., Wu, Z., Hershey, J. R., Saurous, R. A., Weiss, R. J., Jia, Y., and Moreno, I. L. (2019b). VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking. In *Proc. Interspeech 2019*.

Wang, Z.-Q., Le Roux, J., and Hershey, J. R. (2018). Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Wang, Z.-Q. and Wang, D. (2018). Combining spectral and spatial features for deep learning based blind speaker separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(2):457–468.

Wang, Z.-Q., Wang, P., and Wang, D. (2021). Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2001–2014.

Wang, Z.-Q. W., Wang, P., and Wang, D. (2020). Complex spectral mapping for single- and multi-channel speech enhancement and robust asr. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1778–1787.

Warsitz, E. and Haeb-Umbach, R. (2007). Blind acoustic beamforming based on generalized eigenvalue decomposition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 15(5):1529–1539.

Watanabe, S., Mandel, M., Barker, J., and Vincent, E. (2020). CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *ArXiv*, abs/2004.09249.

Weninger, F., Hershey, J. R., Le Roux, J., and Schuller, B. (2014). Discriminatively trained recurrent neural networks for single-channel speech separation. In *2014 IEEE Global Conference on Signal and Information Processing, GlobalSIP, Machine Learning Applications in Speech Processing Symposium*.

Wichern, G., Antognini, J., Flynn, M., Zhu, L. R., McQuinn, E., Crow, D., Manilow, E., and Roux, J. L. (2019). WHAM!: Extending speech separation to noisy environments. In *Proc. Interspeech 2019*.

Williamson, D. S., Wang, Y., and Wang, D. (2016). Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3):483–492.

Wisdom, S., Erdogan, H., Ellis, D., Serizel, R., Turpault, N., Fonseca, E., Salamon, J., Seetharaman, P., and Hershey, J. (2021). What's all the fuss about free universal sound separation data? In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Wisdom, S., Hershey, J. R., Wilson, K., Thorpe, J., Chinen, M., Patton, B., and Saurous, R. A. (2019). Differentiable consistency constraints for improved deep speech enhancement. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R., Wilson, K., and Hershey, J. (2020). Unsupervised sound separation using mixture invariant training. In *Proc. NeurIPS 2020*.

Wolf, M. and Nadeu, C. (2014). Channel selection measures for multi-microphone speech recognition. *Speech Communication*, 57:170–180.

Wölfel, M. (2007). Channel selection by class separability measures for automatic transcriptions on distant microphones. In *Proc. Interspeech 2007*.

Wölfel, M. and Woelfel, M. (2009). *Distant Speech Recognition*, chapter 1, pages 1–25. John Wiley & Sons, Ltd.

Wood, S. U. N., Rouat, J., Dupont, S., and Pironkov, G. (2017). Blind Speech Separation and Enhancement With GCC-NMF. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):745–755.

Wu, J. and Khudanpur, S. (2002). Building a topic-dependent maximum entropy model for very large corpora. In *2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Xiong, F., Zhang, J., Meyer, B., Christensen, H., and Barker, J. (2018). Channel selection using neural network posterior probability for speech recognition with distributed microphone arrays in everyday environments. In *Proc. 5th International Workshop on Speech Processing in Everyday Environments.*

Xu, C., Rao, W., Chng, E., and Li, H. (2020a). Spex: Multi-scale time domain speaker extraction network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1370–1384.

Xu, C., Rao, W., Siong, C. E., and Li, H. (2019a). Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Xu, C., Rao, W., Siong, C. E., and Li, H. (2019b). Time-domain speaker extraction network. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).*

Xu, J., Hu, K., Xu, C., Tran, D. C., and Wang, Z. (2020b). Speaker-aware monaural speech separation. In *Proc. Interspeech 2020.*

Xu, Y., Du, J., Dai, L., 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.

Yoshioka, T., Erdogan, H., Chen, Z., and Alleva, F. (2018a). Multi-microphone neural speech separation for far-field multi-talker speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Yoshioka, T., Erdogan, H., Chen, Z., Xiao, X., and Alleva, F. (2018b). Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks. In *Proc. Interspeech 2018.*

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, D., Kolbæk, M., Tan, Z.-H., and Jensen, J. (2017). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Zeghidour, N. and Grangier, D. (2021). Wavesplit: End-to-end speech separation by speaker clustering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2840–2849.

Zhang, J., Zorilă, C., Doddipatla, R., and Barker, J. (2020a). On end-to-end multi-channel time domain speech separation in reverberant environments. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zhang, J., Zorila, C., Doddipatla, R., and Barker, J. (2021a). Teacher-student mixit for unsupervised and semi-supervised speech separation. In *Proc. Interspeech 2021*.

Zhang, J., Zorila, C., Doddipatla, R., and Barker, J. (2021b). Time-domain speech extraction with spatial information and multi speaker conditioning mechanism. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zhang, J., Zorila, C., Doddipatla, R., and Barker, J. (2022). On monoaural speech enhancement for automatic recognition of real noisy speech using mixture invariant training. *ArXiv*.

Zhang, S., Do, C.-T., Doddipatla, R., and Renals, S. (2020b). Learning noise invariant features through transfer learning for robust end-to-end speech recognition. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zhang, Z., Xu, Y., Yu, M., Zhang, S.-X., Chen, L., and Yu, D. (2021c). ADL-MVDR: All deep learning mvdr beamformer for target speech separation. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zheng, L., Tang, M., Chen, Y., Zhu, G., Wang, J., and Lu, H. (2021). Improving multiple object tracking with single object tracking. *2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Žmolíková, K., Delcroix, M., Kinoshita, K., Ochiai, T., Nakatani, T., Burget, L., and Černockỳ, J. (2019). SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):800–814.

Zorila, C., Li, M., and Doddipatla, R. (2021). An investigation into the multi-channel time domain speaker extraction network. *2021 IEEE Spoken Language Technology Workshop (SLT)*.

Žmolíková, K., Delcroix, M., Kinoshita, K., Higuchi, T., Nakatani, T., and Cernocký, J. H. (2018). Optimization of speaker-aware multichannel speech extraction with asr criterion. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Žmolíková, K., Delcroix, M., Kinoshita, K., Higuchi, T., Ogawa, A., and Nakatani, T. (2017). Speaker-aware neural network based beamformer for speaker extraction in speech mixtures. In *Proc. Interspeech 2017*.