Automatic Speech Separation for Brain-Controlled Hearing Technologies

Cong Han

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2024

Abstract

Automatic Speech Separation for Brain-Controlled Hearing Technologies

Cong Han

Speech perception in crowded acoustic environments is particularly challenging for hearing impaired listeners. While assistive hearing devices can suppress background noises distinct from speech, they struggle to lower interfering speakers without knowing the speaker on which the listener is focusing. The human brain has a remarkable ability to pick out individual voices in a noisy environment like a crowded restaurant or a busy city street. This inspires the brain-controlled hearing technologies. A brain-controlled hearing aid acts as an intelligent filter, reading wearers' brainwaves and enhancing the voice they want to focus on. Two essential elements form the core of brain-controlled hearing aids: automatic speech separation (SS), which isolates individual speakers from mixed audio in an acoustic scene, and auditory attention decoding (AAD) in which the brainwaves of listeners are compared with separated speakers to determine the attended one, which can then be amplified to facilitate hearing. This dissertation focuses on speech separation and its integration with AAD, aiming to propel the evolution of brain-controlled hearing technologies. The goal is to help users to engage in conversations with people around them seamlessly and efficiently.

This dissertation is structured into two parts. The first part focuses on automatic speech separation models, beginning with the introduction of a real-time monaural speech separation model, followed by more advanced real-time binaural speech separation models. The binaural models use both spectral and spatial features to separate speakers and are more robust to noise and reverberation. Beyond performing speech separation, the binaural models preserve the interaural cues of separated sound sources, which is a significant step towards immersive augmented hearing. Additionally, the first part explores using speaker identifications to improve the performance and robustness of models in long-form speech separation. This part also delves into unsupervised learning methods for multi-channel speech separation, aiming to improve the models' ability to generalize to real-world audio. The second part of the dissertation integrates speech separation introduced in the first part with auditory attention decoding (SS-AAD) to develop brain-controlled augmented hearing systems. It is demonstrated that auditory attention decoding with automatically separated speakers is as accurate and fast as using clean speech sounds. Furthermore, to better align the experimental environment of SS-AAD systems with real-life scenarios, the second part introduces a new AAD task that closely simulates real-world complex acoustic settings. The results show that the SS-AAD system is capable of improving speech intelligibility and facilitating tracking of the attended speaker in realistic acoustic environments. Finally, this part presents employing self-supervised learned speech representation in the SS-AAD systems to enhance the neural decoding of attentional selection.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my Ph.D. advisor, Professor Nima Mesgarani. Nima has been a wonderful advisor, giving me invaluable guidance and support throughout my doctoral journey. Nima led me into a fascinating world of computational models and neuroscience, where I had the chance to do unique interdisciplinary research. Nima's passion for impactful research, critical thinking, and long-term vision have continually inspired me. Every time I discussed projects with him, he could point out subtle issues and ask critical questions, and every time I could not answer those questions, I realized that I still had much to improve to become a better researcher. Nima cares about both research development and personal growth of his students. His words have profoundly influenced my life. The five years of my doctoral study under Nima's mentorship have been among the most fulfilling and happiest time of my life.

Second, I express my heartfelt thanks to Yi Luo. Yi initiated my journey to speech separation research. His exceptional talent, keen intuition, and solid research have always impressed me and set an excellent example for me to learn from. Yi has been not only a senior collaborator but also a genuine friend. Sometimes when I had difficulties, I would go to him for a talk. He was always available to listen, share his opinions, and give me encouragement. I was fortunate to have Yi on this journey.

I thank Dr. Zhuo Chen for hosting me at the JSALT workshop in the summer of 2020. Zhuo always sincerely shared his research insights with junior students and provided guidance on their career development, from which I have greatly benefited. My thanks also go to Dr. Simon

Carlile, Dr. Malcolm Slaney, and Dr. Merve Kaya for hosting me at Google X in the summer of 2021. Despite the challenges caused the pandemic, their mentorship and generous support transformed this period into my treasured memory. I am grateful to Simon for offering this incredible opportunity, to Merve for helping me set up everything, and to Malcolm for guiding me to complete a research paper. Additionally, I thank Dr. Kevin Wilson and Dr. John Hershey for hosting me at Google Research in the summer of 2022. Kevin's knowledge, patience, and support were invaluable, and working with John, an icon in speech separation, was a great honor. My thanks extends to many brilliant collaborators and friends I met during these internships: Chenda Li, Keisuke Kinoshita, Marc Delcroix, Mathieu Parvaix, Sheng-Chi Liu, Kyle Hoefer, Scott Wisdom, Hakan Erdogan, Xuankai Chang, Efthymios Tzinis, Qidong Zhao.

I appreciate Professor John Wright, Professor Julia Hirschberg, Professor John Paisley, and Professor Malcolm Slaney for serving on my thesis committee and providing invaluable feedback. I am particularly thankful to Malcolm for his extensive help in my studies. Malcolm gave me countless feedback on the research paper and thesis manuscript and taught me the importance of presenting research work and telling a good story.

I thank all our previous and current Naplab members: James O'Sullivan, Laura Long, Bahar Khalighinejad, Prachi Patel, Zhuo Chen, Yi Luo, Menoua Keshishian, Enea Ceolini, Giovanni Di Liberto, Vinay Raghavan, Vishal Choudhari, Gavin Mischler, Sam Norman-Haignere, Kiki Van der Heijden, Xiaomin He, Yinghao Li, Xilin Jiang, Sukru Samet Dindar, Siavash Shams. They have made my Naplab time even more enriching and memorable. I am thankful to Menoua for his help when I encountered issues accessing the lab servers, and for the numerous enlightening conversations we've had, ranging from research topics to life matters. Vishal has consistently been a supportive, encouraging, and caring friend and collaborator. His attributes of self-discipline, optimism, and a commitment to delivering value have had a high impact on me. I thank Yinghao for teaching me text-to-speech knowledge, and his unique thoughts made every research discussion enjoyable and inspiring. I would also thank Xilin for being a good collaborator and showing me his beautiful code and figures.

Beyond Naplab, my life at Columbia has been enriched by many wonderful friends: Qing Qu, Bowen Yang, Qianyuan Chen, Gaojianyong Wang, Yunzhe Tao, Jingping Nie, Tingkai Liu, Mingyuan Dong, Hongyu Yang, Lei Zuo, Elden Griggs, and many others. The moments we shared are treasured memories I will hold forever. I extend special thanks to Qing Qu, who was my roommate for over two years, including the challenging pandemic period. I am deeply appreciative of his support during my difficult times. Qing's passion and hard work in research were always a source of inspiration to me.

I am grateful to Ms. Marie-Josée Kravis and Mr. Henry R. Kravis for generously providing funding to support part of my research.

Lastly, I would like to give my deepest love and thanks to my parents. Words can never express my gratitude for everything you have done for me.

This is for my Mom and Dad.

Chapter 1: Introduction

1.1 Motivation

Audio and speech technologies have revolutionized the way we interact with the world and each other [1, 2, 3]. Voice assistants like Siri, Alexa, Google Assistant, and ChatGPT voice chat are now a part of our everyday life. They help us manage tasks, answer questions, and control devices in our homes. Speech recognition and speech synthesis technologies have been transformed into a range of applications such as closed captions, speech translation [4], and text-to-speech [5, 6, 7]. Nowadays, it is common to see people wearing noise-cancelling earbuds everywhere. Speech technologies have increased human productivity and improved the quality of our lives.

On the one hand, we enjoy the benefits of technological progress; on the other hand, we have to deal with the complex effects of living in a more advanced world. Today, we are always surrounded by countless sounds, way more than we were decades ago. These sounds range from necessary ones to irrelevant and distracting noises like bustling crowds, operating machinery, and heavy traffic. We can understand and adjust to the acoustic world because our auditory systems can analyze complex sounds, integrate the sound attributes linked to a particular source, and segregate them from those associated with other sources [8]. However, this capacity is limited by strong distractors or listening fatigue. It is even worse for people with hearing disabilities. The next frontier for hearing technology is intelligent hearing assistance devices. They are expected to help people dynamically analyze surrounding acoustic environments and modify or even recreate acoustic scenes to facilitate listening. Imagine we are walking through a busy street, intelligent hearing devices can not only filter out the overwhelming traffic noise to protect our hearing and help us focus on conversations with friends, but also enhance the sounds of approaching vehicles or emergency signals to ensure our safety. In a nature park, intelligent hearing devices can keep the sounds you find pleasant such as bird singing and river flowing, while minimizing the intrusion of unrelated human chatter. In addition, we expect the device to quickly figure out where we are directing our attention and respond better to our specific auditory needs.

As we anticipate the superhuman abilities that intelligent hearing devices may provide in the future, our current priority is to develop a smart hearing aid to address an accessibility challenge faced by many people. In a natural auditory environment, people with normal hearing can effortlessly concentrate on a single speaker among many and switch their attention between speakers seamlessly. However, for those with hearing impairments, this task becomes extremely difficult. People with hearing impairments may need to expend extra listening effort and rely on higher-level compensatory cognitive processes. Traditional hearing aids have seen substantial progress in suppressing background noises that are acoustically different from speech [9, 10], but they cannot enhance a target speaker without knowing which speaker the listener is conversing with [11]. It is therefore important for a hearing aid to be able to automatically distinguish between attended and unattended speakers to selectively enhance the attended speaker's speech.

Fortunately, recent discoveries of the properties of speech representation in the human auditory cortex have shown an enhanced representation of the attended speaker relative to unattended sources [12]. These findings have inspired the idea of auditory attention decoding (AAD) which uses the listener's brain activity to determine which talker the listener is attending to. Knowing the attended speaker, the device can amplify that speaker relative to others to facilitate hearing in a crowd. This provides the solution to creating brain-controlled hearing aids. AAD has been successfully implemented with various ways of acquiring brain activity such as magnetoencephalography (MEG) [13], electroencephalography (EEG) [14], and electrocorticography (ECoG) [12]. However, a critical component for actualizing an AAD system is to get access to individual speakers in mixed audio because the attentional focus of the subject is determined by comparing the brainwaves of the listener with each speaker. In real-world scenarios where only mixed audio is available, the AAD system must be able to separate mixed audio into individual sound sources first. We refer to this as the speech separation problem. After speech separation, the AAD sys-

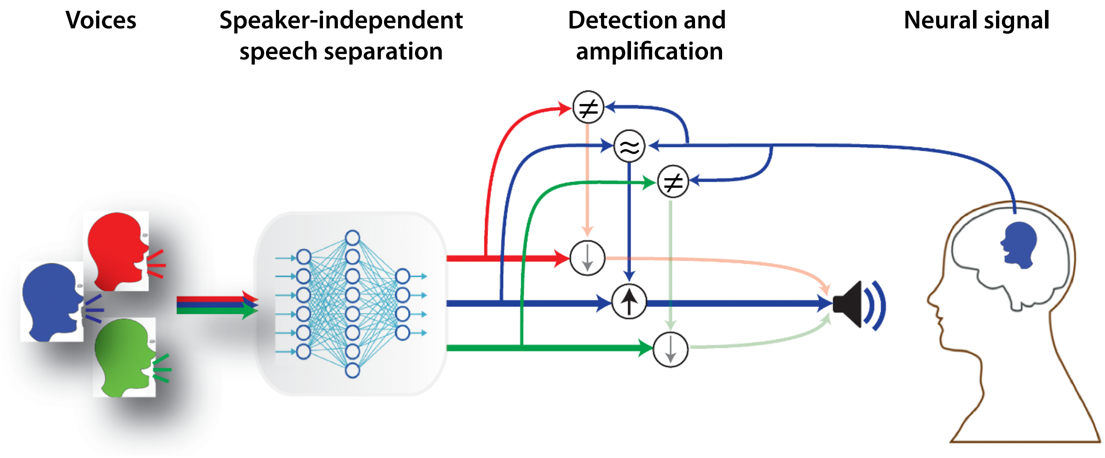

Voices
Speaker-independent speech separation
Detection and amplification
Neural signal

Figure 1.1: **Schematic of the proposed brain-controlled assistive hearing device.** A brain-controlled assistive hearing device can automatically amplify one speaker among many. A deep neural network automatically separates each of the speakers from the mixture and compares each speaker with the neural data from the user's brain to accomplish this goal. Then, the speaker that best matches the neural data is amplified to assist the user.

tem can detect the attended source and subsequently amplify it. Figure. 1.1 shows a schematic of the proposed brain-controlled assistive hearing device that combines speech separation and AAD techniques to selectively amplify the attended speaker.

This dissertation focuses on addressing the speech separation problem to advance brain-controlled hearing technologies. We will use such hearing technologies to help hearing impaired listeners more easily communicate in crowded environments and help these people become more socially engaged and stay connected with family and friends.

1.2 Background

Speech separation has been an important research topic in signal processing for a long time. The main goal is to separate each speaker's voice from mixed sounds. It has benefited a wide range of applications such as telecommunications, automated speech recognition, and the innovation of hearing aids, which we will explore in-depth in this dissertation. Speech separation can be generalized to universal sound separation [15], which aims to separate arbitrary classes of sound from each other. This broader concept is crucial for the future intelligent hearing assistants we discussed earlier, which enable us to understand and adjust the sounds around us. In this dissertation, our emphasis is on speech separation. In this section, we highlight several critical aspects of speech separation, especially for its application in hearing devices.

Speaker-Independent Speech Separation

Before the deep learning era, popular speech separation techniques included non-negative matrix factorization (NMF) [16, 17], probabilistic models [18], and computational auditory scene analysis (CASA) [19]. These methods are speaker-dependent, which means they can be applied to closed-set speakers but not unknown speakers. Brain-controlled hearing aids equipped with these separation models can help a user interact with known speakers, such as family members, but cannot generalize to new, unseen speakers, making it ineffective if the user converses with a new person.

Speaker-independent automatic speech separation means the separation of speakers can be performed without any prior speaker information or training on target speakers, so it can generalize to unseen speakers. Speaker-independent speech separation has been one of the most difficult speech processing problems to solve. Two main contributions to the development of this problem are the deep clustering network (DPCL) and the permutation invariant training method (PIT). DPCL [20, 21] maps the time-frequency (T-F) bins to a high-dimensional embedding space such that each T-F bin is represented by an embedding vector. The training objective is set to minimize a given distance metric of embeddings whose T-F bins belong to the same speaker and maximize that of different speaker's embeddings. After training, traditional clustering algorithms can be applied to the embeddings to calculate the source assignments as the estimated T-F masks. PIT is a general method for any type of objective functions to solve the output permutation problem [22, 23]. It determines the correct output permutation by calculating the lowest value on the selected objective function through all possible output permutations. PIT's effectiveness has significantly accelerated the development of speaker-independent speech separation. It has been a standard choice for separation model training.

Low Latency

Hearing aids have strict requirements for low-latency processing. When a superposition of the original noisy signal and the enhanced signal with delay arrives at the ear drum, it creates unwanted comb filter effects. To avoid this, hearing aids endeavor to introduce a delay of less than 10 ms [24]. The latency of speech separation systems mainly comes from two sources: processing latency and algorithmic latency. Processing latency depends on the model's computational complexity and the hardware's capabilities, while algorithmic latency is determined by the model's configuration. The majority of speech separation models rely on noncausal configurations, which means they need future information to improve speech separation. This need for future data inevitably leads to latency, which is an obstacle to applying these models in hearing devices. For real-time speech separation, models need to be causal, relying only on the current time frame and past information. Common strategies to convert a noncausal configuration into a causal one include but are not limited to modifying the padding in convolutional layers, adjusting attention masks in transformer layers, and changing how statistics of activations are calculated in normalization layers. The model's latency is also highly related to the length of its analysis windows. In time-frequency domain models, a lengthy temporal window, e.g., 32 ms, is often used for short-time Fourier transform (STFT) calculations to ensure adequate frequency resolution, which results in considerable latency. Differently, time-domain models [25] use linear encoders with very short filter lengths to replace STFT. So, time-domain models with causal configurations have lower algorithmic latency and are well-suited for applications requiring real-time processing. In Chapter 2, we will introduce a causal separation model in the time-frequency domain to meet the low-latency requirement. Then, we will also present causal separation models in the time domain with even lower latency.

Preserving Spatial Cues

In real-world multi-talker acoustic environments, humans can easily separate speech and accurately perceive the location of each speaker due to the binaural acoustic features such as interaural time differences (ITDs) and interaural level differences (ILDs). Speech processing methods aimed to modify the acoustic scene are therefore required to not only separate sound sources, but do so in a way to preserve the spatial cues needed for accurate localization of sounds. However, most of the binaural speech separation systems [26, 27, 28] are multi-input-single-output (MISO), and hence lose the interaural cues at the output level which are important for humans to perform sound lateralization and localization [29, 30]. One approach uses head-related transfer function (HRTF) to recreate spatial sound. However, this approach requires robust speaker localization algorithms and either measuring or estimating the HRTF of the listener, adding complexity and making the system dependent on the listener. A more desirable system would directly output stereo sound without requiring additional listener-specific information. Conventional signal processing has explored ways to preserve spatial cues in stereo outputs [31, 32, 33, 34, 35, 36, 37, 38, 39]. As deep learning methods have greatly improved the performance of speech separation, we realized preserving spatial cues had been less studied for deep learning-based models. To address this, we have proposed multi-input-multi-output (MIMO) time-domain speech separation network (TasNet) and optimized the training objective functions to both improve speech signal quality and preserve interaural cues [40]. Additionally, we observed that many separation methods are based on the assumption that sound sources remain stationary, a limitation that restricts their practical application in real-world scenarios. We have addressed this challenge by developing MIMO TasNet on datasets consisting of moving sources [41]. Our deep learning-based model demonstrates superior separation performance and more accurate spatial cue preservation compared to traditional methods. We will introduce them in detail in Chapter 2.

Using Speaker Embedding as Long Contextual Information to Improve Local Separation

A speech signal is typically a very long time-sequence signal. For instance, a 10-second waveform sampled at 16 kHz results in 160,000 time steps, or transformed to thousands of frames in a time-frequency representation. Addressing the challenges of modeling such long sequences is crucial. Various approaches, including stacked dilated convolution layers [25], dual-path RNN architectures [42, 43, 44, 45], and long-range attention [46], have been introduced to handle long sequences in speech separation tasks. However, the role of extended context in speech separation is less obvious than in speech recognition tasks, where longer context provides informative textual information. It is unclear how much contextual information a separation model really needs. Simply increasing the receptive field size of the convolutional models does not guarantee improved performance [25]. Instead, utilizing higher resolution local information, by reducing the stride size (also termed as hop size) of the analysis window, tends to yield more benefits. Therefore, it is important to effectively use long contextual information for speech separation rather than simply enabling models to look at farther and farther distances. We argue that a speaker embedding, such as i-vector [47] or d-vector [48], is a form of long contextual information that can be used to improve speech separation explicitly. Speaker embeddings provide globally consistent information that reflects unique voice characteristics of each speaker, which are helpful for local speech separation. Additionally, since speech separation models often output isolated speaker streams without a predetermined order, speaker embeddings can serve as a reliable cue for tracking individual speakers. There have been numerous studies using speaker embeddings to isolate speakers [49, 50, 51, 52, 48, 53]. However, they require prior information of the target such as a snippet of voice, which contradicts the assumption of speaker-independent speech separation. To address this, we have developed methods that can infer speaker embeddings from mixed audio without extra information and then use speaker embeddings to enhance model performance and robustness [54, 55]. We will introduce these two methods in Chapter 3.

Adapting to Real-World Recordings

Most deep learning-based separation models require supervised training data with input sound mixtures paired with isolated sounds as ground-truth targets. However, recording such pairs of isolated sounds and their mixtures in a real environment is not feasible. Thus for supervised training, input mixtures are constructed by synthetically mixing isolated sound sources. Usually, individual sound sources are also simulated to add environmental effects such as room reverberation. However, models trained on synthetic data can degrade on real recordings due to a mismatch between training and test data. Unsupervised methods overcome these problems by requiring only mixed speech signals. A general category of unsupervised approaches utilizes spatial information to cluster sound sources in space [56, 57, 58, 59]. The posterior cluster labels can be used as masks to isolate the target speech. An approach using the complex angular-central Gaussian mixture model (cACGMM) [57] clusters the signals, and the resulting labels are used as pseudo-target to train a deep clustering model [20]. However, directly applying these unsupervised models on mixed audio can yield undesired performance because the clustering-based methods perform poorly in challenging conditions where spatial features are smeared by room reverberance and strong background noise, especially diffuse noise with no distinct directional features. To mitigate these issues, one strategy is to collect moderately noisy recordings without access to ground-truth signals in real scenarios, which can be well processed by the unsupervised clustering methods. Then, we mix several recordings into a much noisier mixture and take advantage of supervised learning to predict the clean speech signals from the mixture [60]. Another strategy is to design more powerful unsupervised methods that enable the training of separation models directly on real recordings. Training on real recordings can further mitigate the mismatch between synthetic mixed audio and real recordings. Mixture invariant training (MixIT) [61] is a recent unsupervised approach that has demonstrated competitive single-channel sound separation performance. MixIT uses mixtures of mixtures as the "noisy" input and uses the individual mixtures as weak references. The model estimates individual sound sources that can be recombined to reconstruct the original reference mixtures. MixIT has been effective at adapting single-channel [62] and multi-channel [63] speech separation models to real-world meetings. We will introduce both strategies in detail in Chapter 4.

Why Is Speech Separation Necessary in a Brain-Controlled Hearing System?

In this dissertation, we use speech separation models to isolate individual sound sources from mixed audio. Then, by comparing these isolated sound sources with neural activities, we identify and amplify the attended one. A key question arises: Is it necessary to separate every speaker in mixed audio when our ultimate interest lies in the attended speaker? An alternative approach simultaneously processes both mixed audio and brain activities to extract the attended sound source [64, 65, 66, 66]. However, the extensive time and resources required to gather sufficient brain data is a significant bottleneck in training DNN-based brain-informed (alternatively referred to as brain-assisted or neuro-steered) speaker extraction models. Therefore, most models use non-invasive EEG data, which is relatively easier and less expensive to collect compared to invasive neural data. The models mentioned above were trained on non-invasive EEG datasets that contain about 20-30 hours of speech-neural data pairs [67, 68, 69]. Although EEG data can provide enough information needed to decode the attentional focus [14, 70, 71], The SNR of non-invasive EEG is not as high as that of invasive EEG, thus this may come at the expense of reducing the decoding speed of the AAD. While Hosseini et al. [64] showed that their model could track the listener's attention almost instantaneously, there remains uncertainty about the model's efficacy in preserving the high quality of extracted speech while rapidly detecting shifts in attention in single-trial cases.

Rapid advancements in speech BCI research have involved invasive neural recordings [72, 73]. The precision and speed offered by invasive recordings are currently unmatched by non-invasive techniques, making them essential for exploring the upper limits of AAD performance. Our focus in this dissertation is therefore on invasive recordings, aiming to design brain-controlled hearing devices that can quickly and accurately adapt to changes in the listener's attention. However, building a sufficient dataset to train DNN models is particularly challenging for invasive EEG which are costly and difficult to collect. The amount of invasive EEG data we use in this dissertation is less than 30 minutes. As a result, directly training a brain-informed speaker extraction model on this small dataset is suboptimal. How to use large amounts of speech data and tiny amounts of neural data to train a DNN model remains a challenge. Ceolini et al. [74] proposed using a "bridge" feature, such as speech envelope, that can be reconstructed from brain activities and be utilized by the speech extraction model to isolate the target speaker. This approach allows the extraction model to train on arbitrary amounts of audio data, independent of brain signals. A major challenge, however, is the mismatch between the clean envelope used during training and the imperfect envelope reconstructed from the brain during inference. To mitigate this mismatch, Gaussian noise was added to the clean envelope during training. Despite this, the stability of this method does not match the system to be introduced in this dissertation. Another issue is the selection of the appropriate bridge feature. As demonstrated in Chapter 7, speech envelope is not the most effective feature for attentional decoding, leading to further questions about the optimal choice of bridge feature and training of extraction models independent of scarce brain data.

Another notable limitation of these brain-informed speaker extraction models is their "black box" nature, where only input and output are visible and explainable. It is difficult to track how well these models follow the listener's attention in real-time especially when the listener switches attention among speakers. They may extract the unwanted speakers instead of the attended one [75]. In addition, this framework provides less flexibility in controlling the volume of the attended speaker and the other sound sources.

The advantage of performing the speaker separation and speaker selection steps independently is that the method allows us to concentrate exclusively on refining the separation models without the need to model brain signals simultaneously. Therefore, we can use vast amounts of audio data to develop better separation models without being upper bounded by the limited amount of neural data, as shown in Chapter 2, 3, and 4. Simultaneously, the low quantity yet high quality invasive data is sufficient to identify the attended source fast and accurately even in complex acoustic environments, as presented in Chapter 5, 6, and 7. At the intersection of separated speech and neural signal, we can explicitly tell how confident the listener is attending to a particular sound source. We can flexibly adjust the volume of both the attended and unattended speakers and even non-speech sound sources, tailoring the audio output to match the listener's preferences.

1.3 Contribution and Dissertation Outline

This dissertation consists of two parts, exploring speech separation models and their integration with auditory attention decoding to create brain-controlled augmented hearing aids, respectively.

Part I is about pure speech separation without any use of brain data:

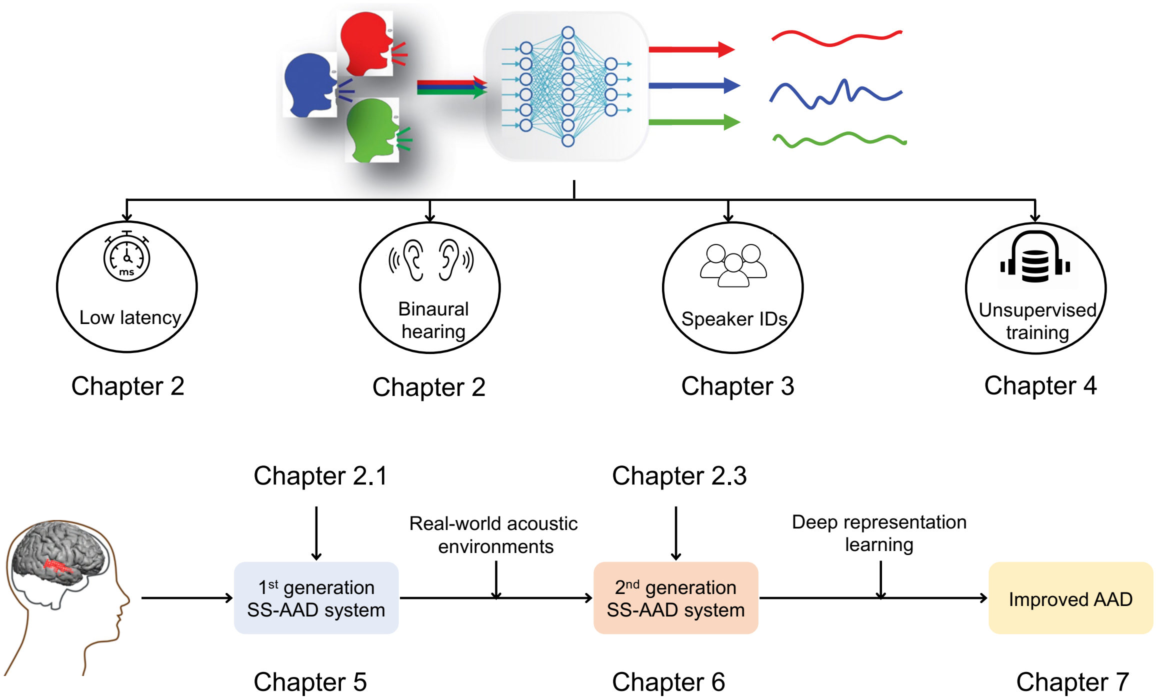


Figure 1.2: Dissertation contributions: Part I, automatic speech separation (top) and Part II, brain-controlled augmented hearing (bottom). SS-AAD means the integration of speech separation and auditory attention decoding.

In Chapter 2, we design a single-channel speech separation model with low latency first. Then, we make a step forward from single-channel speech separation to multi-channel speech separation. We introduce a binaural speech separation model that not only efficiently separates mixed sounds but also accurately preserves interaural cues. This binaural model is designed to enhance immersive listening experiences for hearing aid technologies. The content of this chapter is primarily derived from the previously published work by Han et al. [76, 40, 41].

In Chapter 3, we incorporate speaker information, such as speaker embeddings, into separation models to benefit long-form speech separation. We introduce methods of inferring speaker embeddings from mixed audio directly without requiring extra message and show that using speaker embeddings enhances separation performance and improves speaker tracking capabilities. The content of this chapter is primarily derived from the previously published work by Han et al. [54, 55].

In Chapter 4, we present methods of training multi-channel speech separation models on real-world audio without ground-truth targets. The content of this chapter is primarily derived from the previously published work by Han et al. [60, 63].

Part II introduces brain-controlled augmented hearing systems:

In Chapter 5, we address the problem of speaker-independent AAD without clean sources using the proposed online deep attractor network to automatically separate unseen sources in real time. This chapter proves that integrating speech separation into the AAD framework is a promising approach for developing brain-controlled hearing aids. The content of this chapter is primarily derived from the previously published work by Han et al. [77]. I and James O'Sullivan equally contribute to this chapter.

In Chapter 6, we introduce a realistic AAD experiment paradigm with concurrent conversations and propose the second generation brain-controlled hearing aid that can deal with complex acoustic scene settings with moving talkers and background noise. This chapter represents a substantial leap in applying AAD to real-world settings. The content of this chapter is primarily derived from an unpublished manuscript that was under peer review at the time of writing the dissertation. Vishal Chaudhari and I equally contribute to this chapter.

In Chapter 7, we bring the advancements in self-supervised speech representation learning to the AAD task. This chapter presents a comprehensive demonstration of how self-supervised learned speech representations outperform traditional hand-engineered acoustic features in AAD algorithms. The content of this chapter is primarily derived from the previously published work by Han et al. [78].

At the conclusion of this dissertation, we summarize our key findings and contributions and discuss potential future work in this field.

Part I

Automatic Speech Separation

Chapter 2: Real-Time Speech Separation

Speaker-independent speech separation is a challenging audio processing problem, and recent progress in deep learning has significantly advanced the state of this problem. Deep clustering (DPCL) and permutation invariant training method (PIT) are the two main categories of methods. However, most separation models use noncausal implementation which limits their application in real-time scenarios such as in wearable hearing devices and low-latency telecommunication. In this chapter introduces real-time speech separation models in both categories mentioned above. First, within DPCL, we introduce a real-time monaural speech separation model that utilizes an online clustering strategy in the time-frequency domain. Experimental results that the proposed causal model has competitive performance with other noncausal models employing offline clustering approaches. Second, we propose a real-time binaural speech separation model using PIT in the time domain. The model processes binaural mixed audio, simultaneously separates target speakers in both left and right channels, and maintains the interaural cues of the separated sources. For its application to real-world scenarios with freely moving speakers, we created datasets containing moving sound sources and then developed the model on these datasets. Experimental results show that the proposed binaural model is able to significantly improve the separation performance and keep the perceived location of the modified sources intact in various acoustic scenes.

2.1 Monaural Speech Separation

2.1.1 Introduction

Speaker-independent monaural speech separation is a challenging problem in audio processing. Advancements in deep learning have greatly progressed in addressing this problem. An important example is deep clustering (DPCL) [20, 21] based methods. DPCL maps the time-frequency (T-F) bins to a high-dimensional embedding space such that each T-F bin is represented by an embedding vector. The training objective is set to minimize a given distance metric of embeddings whose T-F bins belong to the same speaker and maximize that of different speaker's embeddings. After training, clustering algorithms can be applied to the embeddings to calculate the source assignments as the estimated T-F masks. Deep Attractor Network (DAN) [79, 80], an extension to DPCL, was proposed to incorporate the clustering step into the network, allowing end-to-end training and evaluation. DAN maps each T-F bin of the mixture spectrogram to a high-dimensional embedding space similar to DPCL, and it explicitly forms clusters by calculating the oracle cluster centers based on the oracle embedding assignment (i.e., ideal T-F mask). The oracle cluster centers are called *attractor points*, and the embeddings that correspond to a specific speaker are constrained to be close to the corresponding attractor point.

However, the successful separation of these models is contingent upon noncausal configuration, which means they require future information from the utterance. This greatly limits the deployment of such systems in real-time applications such as wearable hearing devices. In this section, we propose online DAN (ODAN), a causal extension of the previous noncausal DAN that enables causal and real-time separation. ODAN calculates attractor points for the speakers at each frame, and the sequence of attractor points is tracked with a dynamic weighting function motivated by the online clustering methods [81, 82]. This allows us to perform online clustering and estimate the source assignment frame-by-frame. Experiments show that the proposed ODAN has comparable performance with the noncausal DAN in both two-speaker and three-speaker separation tasks.

2.1.2 Method: Online Deep Attractor Network

Defining the Frequency-Domain Speech Separation

The problem of speech separation is formulated as estimating C sources, $\mathbf{s}_1, \ldots, \mathbf{s}_c \in \mathbb{R}^T$ from the mixture waveform $\mathbf{x} \in \mathbb{R}^T$,

$$\mathbf{x}(t) = \sum_{i=1}^{C} \mathbf{s}_i(t). \tag{2.1}$$

Taking the short-time Fourier transform (STFT) of both sides formulates the source separation problem in the T-F domain where the complex mixture spectrogram is the sum of the complex source spectrograms,

$$\mathbf{X}(f,t) = \sum_{i=1}^{C} \mathbf{S}_i(f,t), \tag{2.2}$$

where $\mathbf{X}$ and $\mathbf{S}_i \in \mathbb{C}^{F \times T}$. One common approach for recovering the individual sources, $\mathbf{S}_i$, is to estimate a real-valued T-F mask for each source, $\mathbf{M} \in \mathbb{R}^{F \times T}$, such that

$$|\hat{\mathbf{S}}_i(f,t)| = |\mathbf{X}(f,t)|\mathbf{M}(f,t). \tag{2.3}$$

The waveforms of the separated sources are then approximated using the inverse STFT of $|\hat{\mathbf{S}}_i|$ using the phase of the mixture audio,

$$\hat{\mathbf{s}}_i = \text{IFFT}\left(|\hat{\mathbf{S}}_i| \odot \angle\hat{\mathbf{X}}_i\right). \tag{2.4}$$

The mask for each source needs to be estimated directly from the mixture spectrogram,

$$\mathbf{M}_i = \mathcal{H}(|\mathbf{X}|; \theta), \tag{2.5}$$

where $\mathcal{H}(\cdot)$ is the mask estimation model defined by parameter θ.

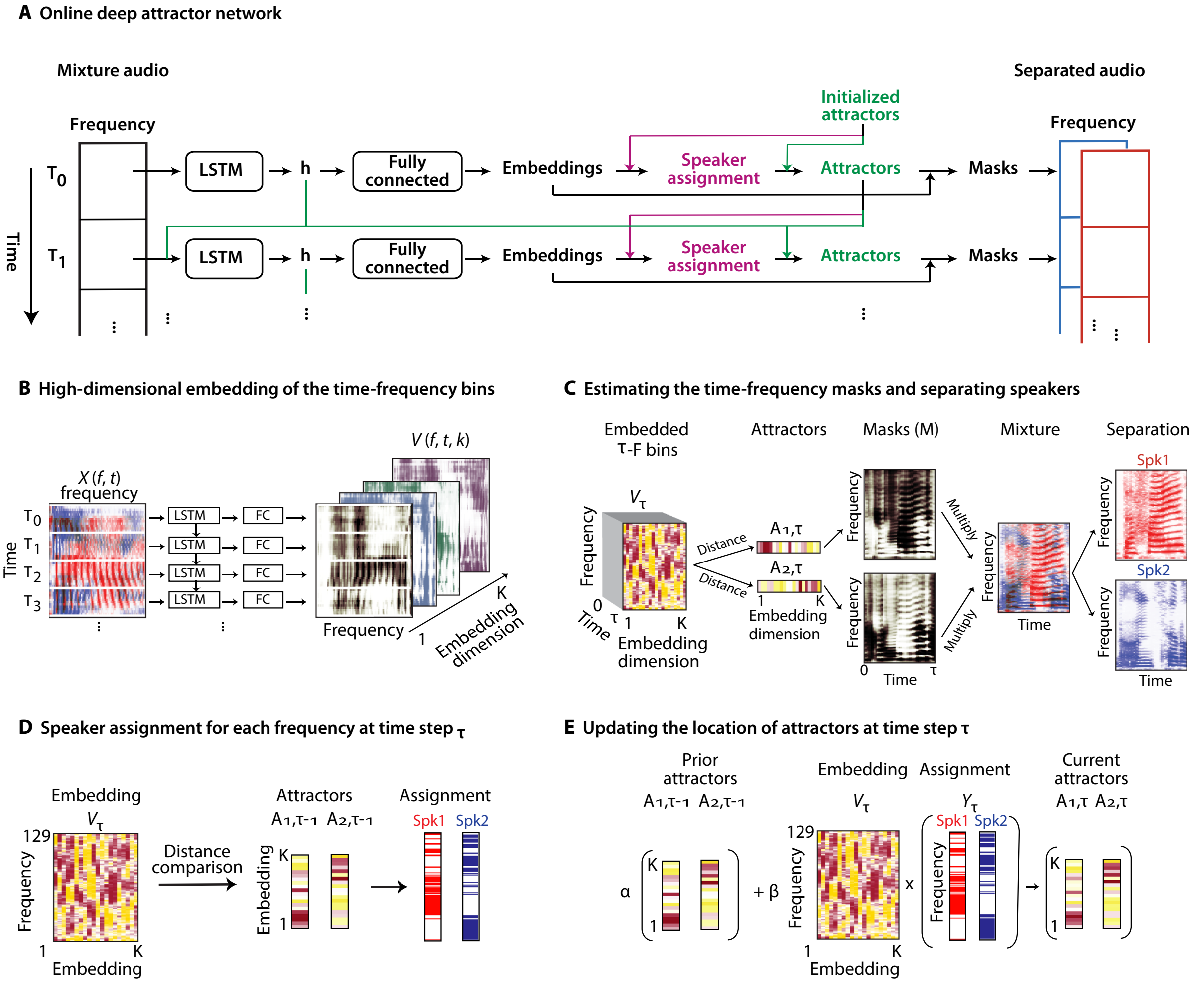


Figure 2.1: **Speaker-independent speech separation with ODAN.** (A) The flowchart of the ODAN for speech separation. (B) The T-F representation of the mixture sound is projected into a high-dimensional space in which the T-F points that belong to the same speaker are clustered together. (C) The center of each speaker representation in the embedding space is referred to as the attractors. The distance between the embedded T-F points and the attractors defines a mask for each speaker that multiplies the T-F representation to extract the speakers. (D) The location of the attractors is updated at each time step. First, the previous location of the attractors is used to determine the speaker assignment for the current frame. (E) Then, the attractors are updated based on a weighted average of the previous attractors and the center of the current frame defined by the speaker assignments.

Online Deep Attractor Network

Figure 2.1A shows the flowchart of the ODAN algorithm. In this novel extension of DAN, source separation is performed by first projecting the mixture spectrogram onto a high-dimensional space where T-F bins belonging to the same source are placed closer together to facilitate their assignment to the corresponding sources. This procedure is performed in multiple steps. First, the mixture magnitude spectrogram, $|\mathbf{X}|$, is projected onto a tensor, $\mathbf{V} \in \mathbb{R}^{F \times T \times K}$, where each T-F bin is represented by a vector of length K (Fig. 2.1B),

$$\mathbf{V} = \Phi(|\mathbf{X}|; \theta), \tag{2.6}$$

where the separation model, $\Phi(\cdot)$, is implemented using a deep neural network with parameter θ. We refer to this representation as the embedding space. The neural network that embeds the spectrogram consists of a four-layer long short-term memory (LSTM) network, followed by a fully connected layer (FC). To assign each embedded T-F bin to one of the speakers in the mixture, we track the centroid of the speakers in the embedding space along time. We refer to the centroids of the source i and at time step τ as the attractor points, $\mathbf{A}_i(\tau) \in \mathbb{R}^K$, because they pull together and attract all the embedded T-F bins that belong to the same source. Therefore, the distance (defined as the dot product [83]) between the embedded T-F bins to each of the attractor points determines the source assignment for that T-F bin, which is then used to construct a mask to recover that source (Fig. 2.1C),

$$\mathbf{M}(f, \tau) = \text{Softmax}\left(\mathbf{A}_i(\tau)\mathbf{V}^{\mathrm{T}}(f, \tau)\right), \tag{2.7}$$

and the Softmax function is defined as

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j^C e^{x_j}}. \tag{2.8}$$

The masks subsequently multiply by the mixture magnitude spectrogram to estimate the magnitude spectrograms of each source (Fig. 2.1C and Eq. 2.3). All the parameters of the ODAN are found jointly during the training phase by minimizing the source reconstruction error of the entire utterance,

$$\mathcal{L} = \sum_{i,f,t} \| |\mathbf{S}_i(f,t)| - |\mathbf{X}(f,t)|\mathbf{M}_i(f,t) \|_2^2. \tag{2.9}$$

Initializing Attractor Points. The initial position of the attractor points at $\tau = 0$ in the embedding space is chosen from a set of N predetermined points, which we refer to as anchor points [80]. During the training phase, we create N randomly initialized, trainable anchor points in the embedding space V, which are denoted by $B_j = 1, \ldots, N$. During the training of the network, the position of the anchor points are jointly optimized to maximize the separability of the mixture sounds. After the training is performed, the anchor points are fixed. To separate a mixture that contains C speakers during the test phase, we first choose all possible C combinations of the N anchor points, resulting in $\binom{N}{C}$ subsets of the N anchors. Next, we find the distance of the embedded T-F bins at $\tau = 0$ from the anchor points in each of the $\binom{N}{C}$ subsets. The C initial attractors for a particular mixture are the ones in the subset that minimize in-set similarity between the attractors (i.e., maximizing the in-set distance between the chosen attractor points).

Online Tracking of the Attractor Points. While DAN uses the embedding of the entire mixture utterance to calculate the attractor points [79, 80], ODAN estimates the attractor locations at each time step using only the current and past inputs. The location of the attractor points in the embedding space is initialized at $\tau = 0$. Updating the attractor points in each time step is performed using a one-step generalized expectation maximization (EM) algorithm [84]. At time step τ, we first calculate the source assignment vectors for each speaker, $\mathbf{Y}_i(f,\tau)$, from the embedded frequency channels $\mathbf{V}(f,\tau)$ by comparing the distance of each embedded T-F bin to each attractor from the previous time step, $\mathbf{A}_i(\tau - 1)$ (Fig. 2.1D),

$$\mathbf{Y}(f,\tau) = \text{Softmax}\left(\mathbf{A}_i(\tau - 1)\mathbf{V}^{\mathrm{T}}(f,\tau)\right). \tag{2.10}$$

A Softmax function is applied to enhance the source assignment contrast. Next, we update the location of the attractors based on the centroid of the current frame, the previous location of the attractors, and the current input (Fig. 2.1E),

$$\mathbf{A}_i(\tau) = (1 - \alpha_i(\tau))\mathbf{A}_i(\tau - 1) + \alpha_i(\tau)\mathbf{C}_i(\tau) \tag{2.11}$$

$$\mathbf{C}_i(\tau, k) = \frac{\sum_f \mathbf{V}(f, \tau, k)\mathbf{Y}_i(f, \tau)}{\sum_f \mathbf{Y}_i(f, \tau)}, \tag{2.12}$$

where $\mathbf{C}_i(\tau) \in \mathbb{R}^K$ is the centroid of the embeddings of source i at time step τ, and parameter α_i determines the rate of the update at time τ by controlling the trade-off between the previous location of the attractors and the centroid of the sources in the current frame. If α is too small, the attractor changes position too quickly from one frame to the next, which may result in a noisy estimate and unstable separation. If α is too large, the attractor will be too slow to track the changes in the mixture condition, which could be problematic if the speakers in the mixture change over time. To optimally estimate α, we calculate a dynamic weighting function to control the relative weight of previous and current estimates using a parameter, $\mathbf{Q}$, for each source i at time step τ.

$$\mathbf{Q}_i(\tau) = \sigma\left(\mathbf{h}(\tau - 1)\mathbf{W} + \mathbf{X}(\tau)\mathbf{U} + \mathbf{A}_i(\tau - 1)\mathbf{J} + \mathbf{b}\right), \tag{2.13}$$

where $\sigma(\cdot)$ is the sigmoid activation function, $\mathbf{h}(\tau - 1)$ is the output of the LSTM layer in the last time step, $\mathbf{X}(\tau)$ is the current mixture feature, and $\mathbf{W},\ \mathbf{U},\ \mathbf{J},\ and\ \mathbf{b}$ are parameters that are jointly learned during the training of the network. Given parameter $\mathbf{Q}_i(\tau)$, the update parameter $\alpha_i(\tau)$ is estimated using the following equation

$$\alpha_i(\tau, k) = \frac{\sum_f \mathbf{Y}_i(f, \tau)}{\mathbf{Q}_i(\tau, k)\sum_{t=0}^{\tau-1} \mathbf{Y}_i(f, t) + \sum_f \mathbf{Y}_i(f, \tau)}, \tag{2.14}$$

where $\mathbf{Q}_i(\tau)$ adjusts the contribution of previous and current attractor estimates at time step τ.

Once the attractors for the current frame are updated, the masks for separating the current frame are calculated using the similarity of the T-F embeddings and each attractor (Fig. 2.1C).

2.1.3 Experimental Settings

ODAN Network Architecture

The network consisted of four unidirectional LSTM layers with 600 units in each layer. The embedding dimension was set to 20 based on the observations reported earlier [80], which resulted in a fully connected layer of 2580 hidden units (20 embedding dimensions times 129 frequency channels) after the LSTM layers. The number of anchors was set to 6 [80]. We trained the models using curriculum training [80], in which we first trained the models on 100-frame-long input segments (0.8 s) and continued training thereafter on 400-frame input segments (3.2 s). The batch size was set to 128. Adam [85] was used as the optimizer with an initial learning rate of 1e-4, which was halved if validation error does not decrease after three epochs. The total number of epochs was set to 150, and early stopping was applied if validation error is not decreased after 10 consecutive epochs. All models were initialized using a pretrained LSTM DAN model. A gradient clip with a maximum norm of 0.5 was applied to accelerate training.

Dataset

The neural network models were trained by mixing speech utterances from the Wall Street Journal corpus [86]. We used the WSJ0-2mix and WSJ0-3mix datasets, which contain 30 hours of training, 10 hours of validation, and 5 hours of test data. The mixed sounds were generated by randomly selecting utterances from different speakers in the WSJ0 training set and mixing them at various signal-to-noise ratios (SNRs), randomly chosen between -2.5 and 2.5 dB. The test set contained 3000 mixtures generated by combining utterances from 16 unseen speakers from the si_dt_05 and si_et_05 subsets. All sounds were resampled to 8 kHz to simplify the models and to reduce computational costs. The input feature is the log magnitude spectrogram computed using a STFT, with 32-ms window length (256 samples) and 8-ms hop size (64 samples), and weighted by the square root of a hamming window. Wiener filter–like masks [20] were used as the training objective.

Table 2.1: Comparison of speech separation accuracy of ODAN with two other methods for separating two-speaker mixtures (WSJ0-2mix dataset) and three-speaker mixtures (WSJ0-3mix dataset). The separation accuracy of ODAN, which is the causal system, is slightly worse but comparable to the other noncausal methods.

Number of Speakers	Method	Causal	SI-SNRi (dB)	SDRi (dB)	PESQ	ESTOI
Two speakers	Original mixture	-	0	0	2.02	0.56
	DAN-LSTM [80]	No	9.1	9.5	2.73	0.77
	uPIT-LSTM [23]	Yes	-	7.0	-	-
	ODAN	Yes	9.0	9.4	2.70	0.77
Three speakers	Original mixture	-	0	0	1.66	0.39
	DAN-LSTM [80]	No	7.0	7.4	2.13	0.56
	uPIT-BLSTM [23]	No	-	0.74	-	-
	DPCL++ [21]	No	7.1	-	-	-
	ODAN	Yes	6.7	7.2	2.03	0.55

Evaluation Metrics

We evaluated and compared the separation performance on the test set using the following metrics: SDR, SI-SNR [80], and PESQ score [87], as well as ESTOI [88] for the evaluation of speech quality and intelligibility.

2.1.4 Results and Discussion

Table 2.1 shows the comparison of the ODAN method with other state-of-the-art speaker-independent speech separation methods on two-speaker and three-speaker mixtures. As seen in Table 2.1, the ODAN method performs well in separating speakers in the mixture and even performs on par with the noncausal DAN method, which computes the separation from the entire utterance using a global clustering of the embeddings.

We also tested the ability of the ODAN in dealing with an unknown number of speakers in the mixture. This was done by assuming the maximum number of speakers to be three and training the algorithm on both two-speaker (WSJ0-2mix) and three-speaker (WSJ0-3mix) datasets. During the test phase, no information about the number of speakers was provided, and the outputs that have low power (less than 20 dB relative to the other outputs) were discarded. As seen in Table 2.2, the

Table 2.2: Speech separation accuracy of ODAN in separating one-, two-, and three-speaker mixtures (WSJ0-mix2 and WSJ0-mix3 datasets). The ODAN was trained on both the WSJ0-mix2 and WSJ0-mix3 datasets and used in all cases.

Number of Speakers	Causal	SI-SNRi (dB)	SDRi (dB)	PESQ	ESTOI
3	Yes	7.0	7.5	2.08	0.56
2	Yes	8.9	9.3	2.63	0.76
1	Yes	**SI-SNR** 24.4	**SDR** 25.0	4.14	0.98

same ODAN network can successfully separate one-, two-, or three-speaker mixtures without any prior information on the number of sources in the mixture during the test phase.

2.2 Binaural Speech Separation with Preserved Spatial Cues

2.2.1 Introduction

Section 2.1, we have introduced a monaural speech separation model. However, monaural signals do not provide directional information. In real-world multi-talker acoustic environments, humans can easily separate speech and accurately perceive the location of each speaker due to the binaural acoustic features such as interaural time differences (ITDs) and interaural level differences (ILDs). Speech processing methods designed to alter the acoustic scene are therefore required to not only separate sound sources but also preserve the spatial cues crucial for accurate localization of sounds.

Additionally, the developments in hardware technologies have made it possible to build binaural hearing devices with microphones placed on both left and right ears that can communicate wirelessly. These technological advances have enabled binaural speech separation algorithms [89, 26, 27, 28] with simultaneous access to both microphone signals. However, most of binaural speech separation systems are multi-input-single-output (MISO), and hence lose the interaural cues at the output level which are important for humans to perform sound lateralization and localization [29, 30]. To achieve binaural speech separation as well as interaural cue preservation, the multi-input-multi-output (MIMO) setting is necessary, and currently, such setting can be divided into three main categories.

The first category of methods add another stage for binaural sound rendering, such as head-related transfer function (HRTF) hypotheses, after a MISO system [36]. This method decouples speech separation and spatial cues preservation, however, it requires robust speaker localization algorithms and a priori knowledge about the HRTF of the listener [90]. Thus, it not only requires additional effort but limits the system to be listener-dependent.

The second category calculates a real-valued spectro-temporal mask and then applies the same mask to both left and right microphone channels [31, 32, 33, 34, 35, 36]. Because both sides obtain the same zero-phase gain, the original interaural cues are preserved. However, the separation performance may be limited because of the single-channel mask estimation and the constraint due to the same gain assumption.

In the third category, complex-valued filters are applied to all available microphone signals simultaneously to generate binaural outputs with an additional constraint on interaural cue preservation. One approach is to use two beamformers at the same time to generate left and right outputs respectively, such as generalized sidelobe canceller (GSC) [37] and binaural minimum variance distortionless response (MVDR) beamformer [38]. Another approach is multi-channel Wiener filter (MWF) [39] that is equivalent to the combination of spatial filtering and spectral post-filtering. There has been a method that exploits the deep neural network to estimate complex ideal ratio masks (cIRM) for both left and right channels [91]. Since these multi-channel methods aim at estimating the desired separated sources in each channel, the spatial information could be naturally preserved.

One common issue for the systems mentioned above is that the system latency can be perceivable by humans, and the delayed playback of the separated speakers might affect the localization of the signals due to the precedence effect [92]. To decrease the system latency while maintaining the separation quality, we can use time-domain speech separation methods network (TasNet) [25]. Different from time-frequency domain models, such as online deep attractor network in Section 2.1, which use lengthy analysis windows in the short-time Fourier Transform (STFT) operation, TasNet uses learnable encoder and decoder with much smaller window size. Recent TasNet-based models have proven their effectiveness in achieving high separation quality and decreasing the system latency [93, 94, 25, 95].

This section looks into multiple methods for formulating TasNet into MIMO systems and investigates their capability of high-quality separation and interaural cue preservation. We propose a MIMO TasNet that takes binaural mixture signals as input and simultaneously separates speech in both channels, then the separated signals can be directly rendered to the listener without post-processing. The MIMO TasNet exploits a parallel encoder to extract cross-channel information for mask estimation and uses mask-and-sum method to perform spatial and spectral filtering for better separation performance. Experiment results show that MIMO TasNet can perform listener-independent speech separation across a wide range of speaker angles and preserve both ITD and ILD features with significantly higher quality than the single-channel baseline. Moreover, the minimum system latency of the systems can be less than 5 ms, showing the potential for real-world implementation in hearable devices.

2.2.2 MIMO TasNet for Binaural Speech Separation

Problem Definition

The problem of binaural speech separation is formulated as the separation of C sources $\mathbf{s}_i^{L,R}(t) \in \mathbb{R}^T$, $i = 1, \ldots, C$ from the binaural mixtures $\mathbf{x}^L(t), \mathbf{x}^R(t) \in \mathbb{R}^T$, where the superscripts L and R denote the left and right channels, respectively. For preserving the interaural cues in the outputs, we consider the case where every single source signal is transformed by a set of head-related impulse response (HRIR) filters for a specific listener:

$$\begin{cases} \mathbf{s}_i^L = \hat{\mathbf{s}}_i \circledast \mathbf{h}_i^L \\ \mathbf{s}_i^R = \hat{\mathbf{s}}_i \circledast \mathbf{h}_i^R \end{cases} \quad i = 1, \ldots, C \tag{2.15}$$

where $\hat{\mathbf{s}}_i \in \mathbb{R}^{T'}$ is the monaural signal of source i, $\mathbf{h}_i^L, \mathbf{h}_i^R \in \mathbb{R}^{T-T'+1}$ are the pair of HRIR filters corresponding to the source i, and $\circledast$ represents the convolution operation. Using the HRIR-transformed signals as the separation targets forces the model to preserve interaural cues introduced by the HRIR filters, and the outputs can be directly rendered to the listener.

MIMO TasNet

TasNet overview. TasNet has been shown to achieve superior separation performance in single-channel mixtures [25]. TasNet contains three modules: a linear encoder first transforms the mixture waveform into a two-dimensional representation similar to spectrograms; a separator estimates C multiplicative functions similar to time-frequency masks based on the 2-D representation; and a linear decoder transforms the C target source representations back to waveforms.

Various approaches have been proposed to extend TasNet into the multi-channel framework [96, 97]. A standard pipeline is to incorporate cross-channel features into the single-channel model, where spatial features such as interaural phase difference (IPD) are concatenated with the mixture encoder output on a selected reference microphone for mask estimation [96]. In various scenarios, such configuration has led to a significantly better separation performance than the signal-channel TasNet.

Feature concatenation for MIMO TasNet. Recent studies have proposed several approaches to allow TasNet to take multi-channel inputs. A common approach is the integration of cross-channel features into single-channel models. Gu et al. [96] concatenated IPD feature with the encoder output at a selected reference microphone to perform single-channel separation, where the IPD between the complex-valued spectrograms of two signals $\mathbf{X}, \mathbf{Y} \in \mathbb{C}^{F \times T}$ is defined as:

$$\mathrm{IPD}(\mathbf{X}, \mathbf{Y}) = \angle\mathbf{X} - \angle\mathbf{Y} \tag{2.16}$$

where $\angle(\cdot)$ denotes the element-wise function for calculating the angle of a complex number. Practically the IPD features transformed by cosine or sine functions are used for better performance

[96]. As IPD is a T-F domain feature, the window size for short-time Fourier transform (STFT) is typically much larger than that of TasNet. We calculated the IPD feature in T-F domain using a context window whose center is the corresponding window for TasNet. We kept both frequency resolution and time resolution of T-F domain features for separation only in the center window.

Different from multi-channel signals received by the far-field microphone array whose inter-channel level differences can be ignored, binaural signals contain distinct interaural level differences (ILDs), where ILD between the complex-valued spectrograms of two signals $\mathbf{X}, \mathbf{Y} \in \mathbb{C}^{F \times T}$ is defined as:

$$\mathrm{ILD}(\mathbf{X}, \mathbf{Y}) = 10 \log_{10} \left(\frac{|\mathbf{X}|}{|\mathbf{Y}|} \right) \tag{2.17}$$

so, ILD is also concatenated with IPD and the encoder output along the feature dimension. To adapt to a MIMO system, each of the left and right channel signals will be selected as the reference channel and performed separation. For example, given $\mathbf{X}, \mathbf{Y}$ as left and right channel mixtures, we choose $\mathbf{X}$ as the reference and concatenate $\mathrm{IPD}(\mathbf{X}, \mathbf{Y})$ and $\mathrm{ILD}(\mathbf{X}, \mathbf{Y})$ with the encoder output in the single-channel model to separate $\mathbf{X}$, meanwhile we take $\mathbf{Y}$ as the reference and add $\mathrm{IPD}(\mathbf{Y}, \mathbf{X})$, $\mathrm{ILD}(\mathbf{Y}, \mathbf{X})$ to separate $\mathbf{Y}$.

Design of MIMO TasNet. The proposed MIMO TasNet uses a parallel encoder for spectro-temporal and spatial features extraction and a mask-and-sum mechanism for source separation. A *primary* encoder is always applied to the channel to be separated, and a *secondary* encoder is applied to the other channel to jointly extract cross-channel features. In other words, the sequential order of the encoders determines which channel (left of right) the separated outputs belong to. The outputs of the two encoders are concatenated and passed to the separator, and $2C$ multiplicative functions are estimated for the C target speakers. C multiplicative functions are applied to the *primary* encoder output while the other C multiplicative functions are applied to the *secondary* encoder output, and the two multiplied results are then summed to create representations for C separated sources. We denote it as the *mask-and-sum* mechanism to distinguish it from the other

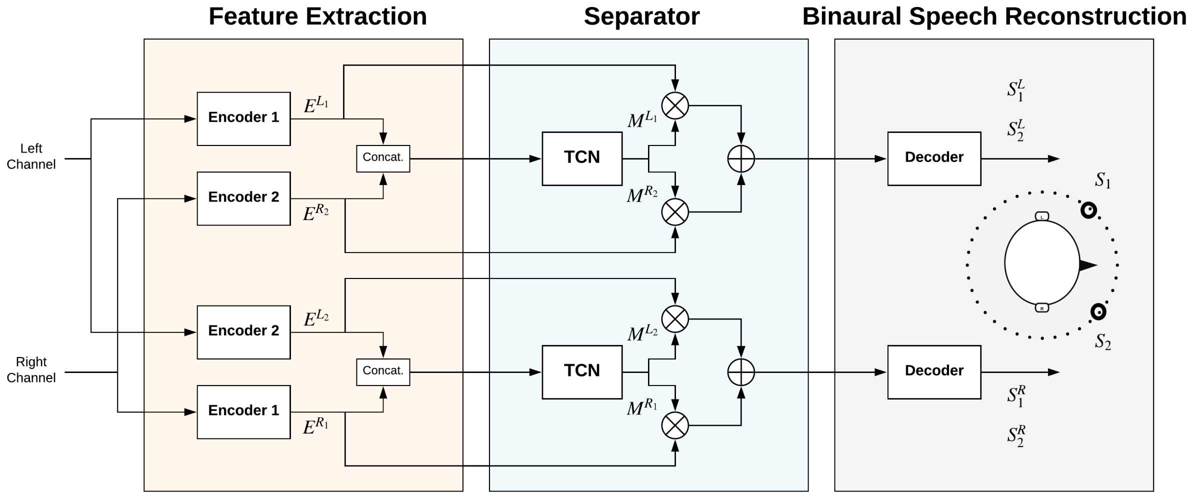


Figure 2.2: **The architecture of the proposed binaural speech separation network.** Two encoders are shared by the mixture signals from both channels, and the encoder outputs for each channel are concatenated together and passed to a mask estimation network. Then, spectral-temporal and spatial filtering are performed by applying the masks to the corresponding encoder outputs and sum them up on both left and right paths. Finally, binaural separated speech are reconstructed by a linear decoder.

methods where only C multiplicative functions were estimated from the separation module and applied to only the reference channel. Similar to TasNet, a linear decoder transforms the C target source representations back to waveforms. Figure 2.2 shows the flowchart of the system design.

Note that a parallel encoder design for multi-channel TasNet has been discussed in a previous literature [96]. For a N-channel input, N encoders are applied to each of them and the encoder outputs are summed to create a single representation. The multiplicative function is also estimated on the single representation, which results in a MISO system design. We can easily find that it is a special case of MIMO TasNet where the two multiplicative functions for the two encoders are equal. Although a previous study found that the feature concatenation method performed comparably to the parallel encoder design, we will show that MIMO TasNet is able to significantly surpass feature concatenation TasNet in various configurations in both separation performance and spatial cue preservation accuracy in Section 2.2.4.

Training objective. Section 2.1 has introduced the deep clustering approach for tackling the speaker-independent speech separation challenge. Another very important method that has significantly advanced speech separation is permutation invariant training (PIT). PIT is a general method for any type of objective functions to solve the output permutation problem [22, 23]. It determines the correct output permutation by calculating the lowest value on the selected objective function through all possible output permutations. Variants on the network architecture design and objective function design have proven the effectiveness of this training method [25]. Due to its broad applicability and effectiveness, PIT has become the most favored training method in speaker-independent speech separation. In this section, we used utterance-level permutation invariant training (u-PIT) [23],

$$\mathcal{L} = \min_{\pi \in P} \sum_{c=1}^{C} \psi(\hat{\mathbf{s}}_c^L, \mathbf{s}_{\pi(c)}^L) + \psi(\hat{\mathbf{s}}_c^R, \mathbf{s}_{\pi(c)}^R) \tag{2.18}$$

where $\psi(\cdot)$ is the objective function between estimated signals and target signals, $\hat{\mathbf{s}}^L, \hat{\mathbf{s}}^R \in \mathbb{R}^T$ are the separated signals in left and right channels, respectively, $\mathbf{s}^L, \mathbf{s}^R \in \mathbb{R}^T$ are the corresponding target signals, the subscript c is the speaker index, and P is the set of all C! permutations.

Most separation models use scale-invariant signal-to-distortion ratio (SI-SDR) as both the evaluation metric and objective function ($\psi(\cdot)$). SI-SDR between a signal $\mathbf{s} \in \mathbb{R}^T$ and its estimate $\hat{\mathbf{s}} \in \mathbb{R}^T$ is defined as:

$$\text{SI-SDR}(\mathbf{s}, \hat{\mathbf{s}}) = 10 \log_{10} \left(\frac{||\alpha \mathbf{s}||_2^2}{||\hat{\mathbf{s}} - \alpha \mathbf{s}||_2^2} \right) \tag{2.19}$$

where $\alpha = \hat{\mathbf{s}}\mathbf{s}^\top / \mathbf{s}\mathbf{s}^\top$ corresponds to the rescaling factor. Although SI-SDR is sensitive to time shift of the estimated signal and thus able to implicitly preserve the ITD information, the scale-invariance property of SI-SDR makes it insensitive to power rescaling of the estimated signal, which may fail in preserving the ILD between the outputs. Hence instead of using SI-SDR as the training objective, we used the plain signal-to-noise ratio (SNR), defined as:

$$\mathrm{SNR}(\mathbf{s}, \hat{\mathbf{s}}) = 10 \log_{10} \left(\frac{||\mathbf{s}||_2^2}{||\hat{\mathbf{s}} - \mathbf{s}||_2^2} \right). \tag{2.20}$$

2.2.3 Experimental Settings

Datasets

We generated an anechoic speech dataset from the WSJ0-2mix dataset [20]. 30 hours of training data, 10 hours of validation data and 5 hours of test data were generated with the same configuration as the single-channel WSJ0-2mix data, while the clean speech was convolved with randomly sampled HRIR filters from the CIPIC HRTF Database [90]. The CIPIC HRTF Database contains real-recorded HRIR filters for 20 subjects across 25 different interaural-polar azimuths from -80° to 80° and 50 different interaural-polar elevations from -90° to 270°. Two speaker locations were randomly sampled from the database for spatial rendering. We used 27 subjects for the training and validation sets and 9 unseen subjects for the test set, ensuring that the model was evaluated in a listener-independent way. All mixtures were downsampled to 8k Hz.

The anechoic WSJ0-3mix dataset with spatial cues was generated by using the same method as above. To generate the noisy WSJ0-2mix dataset, we added to the training set the noise from one out of eight environmental noises (washing room, kitchen, sport field, city park, office, meeting room) chosen from DEMAND dataset [98] with SNR between -2.5 and 15 dB. The noise in the test set was from another eight scenarios (subway station, restaurant, public square, traffic intersection, subway, private car). To generate the echoic WSJ0-2mix dataset, we obtained HRIR filters from the BRIR Sim Set[1], which was simulated with different reverberation time (T60). We added reverberation using rooms with T60 0.1s, 0.2s, 0.4s, 0.5s, 0.7s, 0.8s, 1.0s for training and 0.3s, 0.6s, 0.9s for testing.

[1] `http://iosr.uk/software/index.php`

Evaluation Metrics

We evaluated the model with both the separation quality and the ability to preserve interaural cues. SNR improvement (SNRi) was used as the signal quality metric instead of SI-SDR improvement according to our discussion in Section 2.2.2. ITD and ILD errors between the separated and target clean signals were used as the metric for the accuracy of preserving interaural cues, which are defined as:

$$\Delta_{ITD} = \left|\mathrm{ITD}(\mathbf{s}^{\mathrm{L}}, \mathbf{s}^{\mathrm{R}}) - \mathrm{ITD}(\hat{\mathbf{s}}^{\mathrm{L}}, \hat{\mathbf{s}}^{\mathrm{R}})\right| \tag{2.21}$$

$$\Delta_{ILD} = \left|10\log_{10}\frac{||\mathbf{s}^{L}||_2^2}{||\mathbf{s}^{R}||_2^2} - 10\log_{10}\frac{||\hat{\mathbf{s}}^{L}||_2^2}{||\hat{\mathbf{s}}^{R}||_2^2}\right| \tag{2.22}$$

where $||\cdot||$ denotes the L_2-norm of the signal. We used generalized cross-correlation phase transform (GCC-PHAT) algorithm [99] to compute time difference of arrival (TDOA) of $\mathbf{s}^L$ and $\mathbf{s}^R$ as $\mathrm{ITD}(\mathbf{s}^{\mathrm{L}}, \mathbf{s}^{\mathrm{R}})$. The tool is available online[2].

Network Architectures

The configurations of the MIMO TasNet variants were based on the causal setting of the single-channel TasNet [25]. In the linear encoder and decoder, we utilized 64 filters, each with a filter length of 2 ms, which equates to 16 samples at a sampling rate of 8 kHz. In the causal temporal convolutional network (TCN), there were 4 repeated stacks and each one included 8 1-D convolutional blocks. The number of parameters in all models was fixed at 1.67M for a fair comparison.

For baseline models, we adopt the following configurations:

1. *Single-channel TasNet*: the single-channel model is applied to each channel independently.

2. *Feature concatenation TasNet*: cross-channel features are concatenated to the encoder output in the same way as [96]. Spatial features used include sin(IPD), cos(IPD) and ILD, where

[2] https://www.mathworks.com/help/phased/ref/gccphat.html

the IPD and ILD are defined as

$$\mathrm{IPD}(\mathbf{X},\mathbf{Y}) = \angle\mathbf{X} - \angle\mathbf{Y} \tag{2.23}$$

$$\mathrm{ILD}(\mathbf{X},\mathbf{Y}) = 10\log_{10}\left(|\mathbf{X}| \oslash |\mathbf{Y}|\right) \tag{2.24}$$

where $\mathbf{X}, \mathbf{Y}$ are the spectrograms of the two channel mixtures, $\oslash$ means element-wise division. The window length of STFT for calculating spectrograms is 256 samples.

3. *Parallel encoder TasNet*: the same configuration as in [96] which is also discussed in Section 2.2.2.

2.2.4 Results and Discussion

Table 2.3 compares different MIMO TasNet variants at various speaker locations on anechoic spatialized WSJ0-2mix. The single-channel baseline is able to achieve the smallest ILD error across all models when the speaker angle is very small, which indicates that the interaural features in this scenario are not helpful in preserving the absolute energy of the separated speech. For all other speaker locations, both the ILD error and separation quality for the single-channel model are significantly worse than all the MIMO variants. For TasNet concatenated with sin(IPD), cos(IPD) and ILD features, we can observe significant signal quality improvement and ITD/ILD error reduction across all angle ranges, and better performance is achieved with larger speaker angle. This confirms the previous observations regarding the effectiveness of cross-channel features in end-to-end frameworks [96]. The parallel encoder method has on par performance in preserving ITD/ILD with feature concatenation method, but achieves better separation performance except when the speaker angle is small (less than 15°). The significant improvement for signal quality (SNRi) indicates that the parallel encoders are able to implicitly extract more effective cross-channel features than cross-domain features IPD/ILD for multi-channel speech separation. The further improvement from mask-and-sum mechanism indicates the effectiveness of combining spatial filtering and spectral filtering to separate sources. The correlation (Pearson's r) between SNRi and Δ_{ITD} and be-

Table 2.3: SNR improvement (dB), ITD error (μs), and ILD error (dB) for different variants of TasNet on anechoic spatialized WSJ0-2mix. The averaged performance on different ranges of speaker angles is reported.

Method	SNRi / Δ_{ITD} / Δ_{ILD}				
	Angle				
	<15°	15-45°	45-90°	>90°	Average
TasNet	10.0 / 6.0 / **0.29**	10.0 / 5.8 / 0.39	10.3 / 4.8 / 0.56	10.7 / 6.4 / 0.59	10.2 / 5.8 / 0.46
+ILD	11.1 / 3.1 / 0.31	13.4 / 1.9 / 0.12	14.4 / 1.3 / 0.16	14.8 / 1.9 / 0.17	13.4 / 2.0 / **0.19**
+sin(IPD), cos(IPD)	11.7 / **2.4** / 0.34	14.1 / 1.7 / 0.14	14.7 / 1.4 / 0.20	15.3 / 2.0 / 0.20	13.9 / 1.9 / 0.22
+sin(IPD), cos(IPD), ILD	**11.8** / **2.4** / 0.33	14.5 / 1.6 / **0.11**	15.3 / 1.2 / 0.16	15.8 / **1.8** / 0.18	14.4 / **1.8** / 0.20
+parallel encoder	10.6 / 3.0 / 0.47	15.1 / 1.5 / **0.11**	16.8 / 1.2 / 0.11	17.7 / 2.0 / 0.12	15.0 / 2.0 / 0.20
+parallel encoder, mask&sum	10.7 / 2.8 / 0.47	**15.6** / **1.3** / 0.13	**17.7** / **1.1** / **0.09**	**18.3** / **1.8** / **0.09**	**15.6** / **1.8** / **0.19**

tween SNRi and Δ_{ILD} are -0.77 and -0.85, respectively (p <0.0001 for both), which means higher separated signal quality helps in preserving ITD/ILD better.

To further examine our proposed MIMO TasNet in more adverse environments, we tested the separation accuracy in three speaker mixtures, noisy speech separation and speech separation with room reverberation. Note that in the evaluation of these three cases, the top 5% Δ_{ITD} and Δ_{ILD} were dropped before averaging to prevent the errors incurred by outliers.

When testing the model on the noisy WSJ0-2mix dataset, we set the noise power range at three levels. As shown in Table 2.4, additive noise contaminates both speech quality and ITD/ILD preservation, but the overall performance compared to the clean condition is still superior and MIMO TasNet with parallel encoder and mask-and-sum achieves the best performance in all metrics across all noise levels, which proves the MIMO TasNet is more robust to the noise.

We observe that three-speaker separation is more challenging than noisy speech separation. Both ITD and ILD preservation downgrade significantly compared to the two-speaker case. That is because the model had failed to separate some of speech with very small power compared to the other two speakers or speech with very similar spatial features to others, and the failure of separation leads to the failure of ITD/ILD preservation.

Finally, we evaluated the model on the echoic spatialized WSJ0-2mix dataset. The target is the reverberant clean signal. Not surprisingly, convolutive room reverberation is a more challenging condition than additive environmental noises in terms of both signal quality improvement and pre-

Table 2.4: Evaluation of TasNet with parallel encoder on several adverse conditions: three-speaker separation, two-speaker separation with environmental noise, and with room reverberance.

Method	SNRi / Δ_{ITD} / Δ_{ILD}						
	3 speaker	2 speaker with noise (SNR)			2 speaker with reverberance (RT60)		
		12.5 dB	5 dB	-2.5 dB	0.3s	0.6s	0.9s
TasNet	9.1 / 6.3 / 0.74	9.8 / 3.4 / 0.31	10.9 / 3.7 / 0.31	13.8 / 5.2 / 0.57	7.2 / 10.8 / 0.46	6.2 / 45.1 / 0.47	5.7 / 44.5 / 0.50
+parallel encoder	11.3 / 12.3 / 0.84	13.7 / 2.3 / 0.16	15.0 / 2.5 / 0.18	17.8 / 3.0 / 0.23	9.2 / 6.5 / **0.20**	7.7 / 33.2 / 0.25	6.9 / 17.7 / 0.30
+parallel encoder, mask&sum	**12.1 / 5.7 / 0.45**	**14.3 / 2.2 / 0.14**	**15.3 / 2.3 / 0.15**	**18.2 / 2.8 / 0.21**	**9.4 / 5.9** / 0.23	**7.8 / 30.0 / 0.21**	**7.1 / 15.6 / 0.25**

serving spatial cues as the sparseness properties of the speech are affected by room reverberation. The smearing caused by reverberation means that the mixture at each instance includes components of the same and different speakers, which makes the mask prediction and TDOA estimation more difficult. As a result, SNRi and Δ_{ITD} are more easily affected by the reverberation. Also, using only two channels does not fully take advantage of multi-channel algorithms to reduce the influence of reverberation. Nonetheless, the averaged 9.4 dB SNR improvement, 5.9 μs ITD error and 0.23 dB ILD error show that the performance of MIMO TasNet is still helpful in the moderate reverberant environment.

2.3 Binaural Speech Separation of Moving Speakers

2.3.1 Introduction

In Section 2.2, we have introduced a real-time binaural speech separation model that can preserve the spatial cues such as interaural level difference (ILD) and the interaural time difference (ITD) of all directional sources to enable a listener to perceive the correct location of sources in space. Similar to most conventional speech separation models, it assumes the sources are not moving, which limits its application in real-world scenarios. A commonly used solution to addressing moving sources is block-wise adaptation of models, meaning to split the signal into very short blocks in which the sources within each block are assumed stationary since spatial features vary smoothly and slowly [100, 101, 102]. Choosing an appropriate block size is not trivial: on one hand, the block size must be small enough to satisfy the stationarity assumption; on the other hand, the block size should be long enough for successful separation. For example, independent component analysis (ICA) methods require a block size that is large enough for the independence assumption to hold within the block. Besides, block-wise methods require a tracking module to resolve the source permutation problem across consecutive blocks [103, 104].

An alternative solution is to divide the moving source separation problem into speech source localization, tracking, and separation problems, and to tackle them separately [105, 106] or jointly [107]. Taseska et al. [108] used an approximate Bayesian tracker that employs the Markovian property of the speaker motion model to associate time-frequency bins to each source, based on which spatial filters are estimated to separate moving sources. However, these approaches depend on the high fidelity of source localization and/or tracking. In real scenarios, speech turns, speaker appearance and disappearance, and room reverberation [109] complicate the source tracking considerably. Moreover, the tracking methods mainly utilize the past spatial information but do not take advantage of long-term spectral information. Long-term spectral information has proven beneficial for source separation [25, 44]. Instead, a method that is able to use both spectral-temporal and spatial-temporal information in a larger context is desired for resolving these challenges.

This section aims to enhance the binaural speech separation model to handle scenarios with moving speakers effectively. To achieve this, we developed a dataset that replicates the dynamics of moving speakers in reverberant environments. Training the model on this dataset allows it to utilize extended contextual information at the utterance level, thereby improving its ability to separate speech and track speaker movements within an utterance. This approach eliminates the need for extra localization and tracking modules. Experimental results show that utterance-level separation significantly outperforms the block-wise adaptation methods in terms of both signal quality and spatial cue preservation.

2.3.2 Method

Figure 2.3 shows the overall flowchart of the proposed framework. In the first step, the binaural speech separation module simultaneously separates the speaker in each channel of the mixed input. In the second step, the binaural post enhancement module enhances each speaker individually.

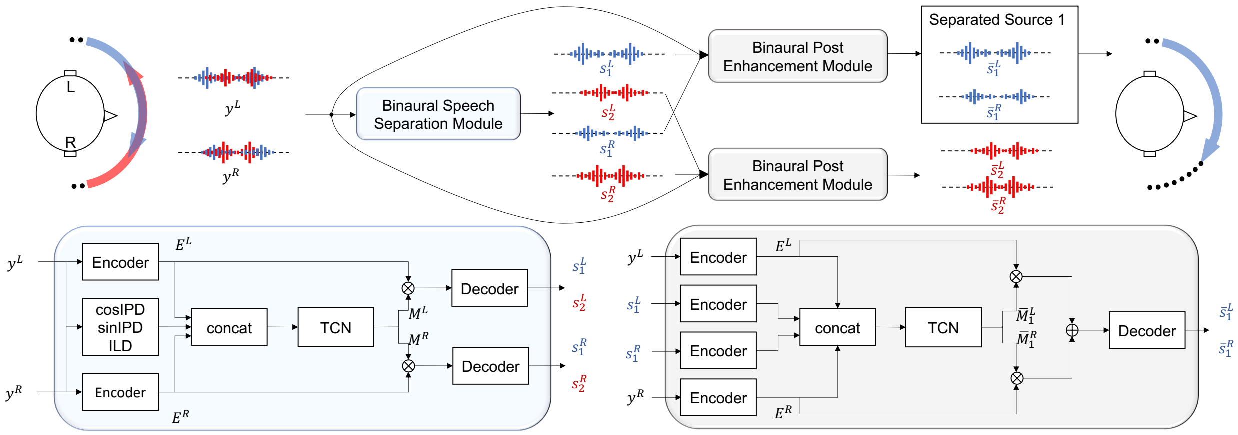


Figure 2.3: **The architecture of the proposed system.** Two moving speakers, denoted in blue and red are being separated. The bottom left and bottom right figures illustrate the details of binaural speech separation and binauarl post enhancement modules.

The binaural speech separation module presented in this section is almost identical to the one described in Section 2.2.2, with a key distinction: here, the module concurrently separates speakers in both the left and right channels, rather than processing the two channels independently. There is an additional binaural post enhancement module to enhance the model's resilience against background noise and reverberation in difficult environments. While this module may introduce some latency, the TasNet architecture effectively maintains acceptable overall latency. To evaluate the effectiveness of our model in preserving interaural cues in the stereo output, a speaker localizer is employed to test whether the separated speakers can be accurately localized.

Binaural Post Enhancement Module

Each stereo sound, $\mathbf{s}_i^L$ and $\mathbf{s}_i^R$, from the separation module, combined with the mixed signals ($\mathbf{y}^L$, $\mathbf{y}^R$), is sent to a multi-input-single-output (MISO) TasNet for post enhancement. Similar to the speech separation module, we concatena all the encoder outputs and pass them to the TCN blocks for estimating multiplicative functions $\mathbf{M}_i^L, \mathbf{M}_i^R \in \mathbb{R}^{2\times N\times H}$,

$$\mathbf{s}_i^L = \text{decoder}(\mathbf{E}^L \cdot \mathbf{M}_i^L[0,:,:] + \mathbf{E}^R \odot \mathbf{M}_i^L[0,:,:]) \tag{2.25}$$

$$\mathbf{s}_i^R = \text{decoder}(\mathbf{E}^L \cdot \mathbf{M}_i^L[1,:,:] + \mathbf{E}^R \odot \mathbf{M}_i^L[1,:,:]) \tag{2.26}$$

where, $\odot$ denotes element-wise multiplication. Different from the speech separation module that only applies multiplicative functions, which is equivalent to spectral filtering, speech enhancement module performs multiplication and sum, which is equivalent to both spectral and spatial filtering. This is similar to multichannel Wiener filtering [39]. We denote it as mask-and-sum.

Since the input stereo sound, $\mathbf{s}_i^L$, $\mathbf{s}_i^R$, contains both spectral and spatial information of the speaker i, the enhancement module essentially performs informed speaker extraction without the need for permutation invariant training.

Speaker Localizer

The speaker localizer adopts a similar architecture as the speech enhancement module but performs classification of the direction of arrival (DOA). We discretize the DOA angles into K classes. The speaker localizer takes only stereo sound, $\mathbf{s}_i^L$, $\mathbf{s}_i^R$, as input, concatenates two encoders' outputs, and passes them to the TCN blocks to estimate a single-class classification matrix $\mathbf{V}_i \in (0, 1)^{K \times H}$, where "single-class" means that in each time frame, there is exactly one class labeled with 1 and all the other classes are labeled with 0.

We split $\mathbf{V}_i$ into B small chunks $\left\{\mathbf{V}_i^b\right\}_{b=1}^{B} \in \mathbb{R}^{K \times Q}$, where Q is the number of time frames in each chunk and $B = \frac{H}{Q}$. In each chunk, we count the frequency of each class labeled with 1, and regard the most frequent class as the estimated DOA for that chunk.

Training Objective

The training objective for the speech separation and enhancement modules is SNR, which is sensitive to both time shift and power scale of the estimated waveform, so it is able to force the ITD and IPD to be preserved in the estimated waveform. In the speech separation module, we used utterance-level permutation invariant training [23] (see Section 2.2.2). To minimize utterance-level separation error, the model must accurately separate and track speakers throughout the utterance. Therefore, the model learns to utilize a significantly longer contextual window than a small block, enabling effective speaker separation and tracking, even in cases of speaker movement.

2.3.3 Experimental Settings

Binaural Room Impulse Responses and Speech Data

We used two types of binaural room impulse responses (BRIRs): one obtained from simulated rooms and the other was measured in real rooms[3]. There were 11 simulated rooms with reverberation time (RT60) varying from 0 to 1 s with 0.1 s increments. In this study, only 8 rooms with RT60 from 0 to 0.7 s were used. There were 4 real rooms with RT60 0.32 s, 0.47 s, 0.68 s, 0.89 s, respectively. The impulse responses were calculated with the sound source located on the frontal azimuthal plane between −90° and 90° with 5° increments at a distance of 1.5 m to the receiver. Two speakers were randomly selected from the 100-hour Librispeech dataset [110]. Both speech data and BRIRs were sampled at 16 kHz.

Moving Source Simulation

Given a monaural speech **s** and a set of BRIRs $\left\{h_j^L\right\}_{j=1}^N, \left\{h_j^R\right\}_{j=1}^N \in \mathbb{R}^{L_h}$, where h_j^L and h_j^R are the BRIR filters of length L_h from the location j to the left ear and right ear, respectively, and N is the number of locations (37 in this study), the moving binaural source was simulated as:

$$\mathbf{s}^L[n] = \sum_{j=0}^{N} \sum_{k=0}^{L_h} \mathbb{I}_j(n) \cdot h_j^L[k] \cdot s[n-k] \tag{2.27}$$

$$\mathbf{s}^R[n] = \sum_{j=0}^{N} \sum_{k=0}^{L_h} \mathbb{I}_j(n) \cdot h_j^R[k] \cdot s[n-k] \tag{2.28}$$

where $\mathbb{I}_j(n)$ is an indicator function which is 1 when **s** is at location j at time step n, and is 0 otherwise. This method simulates the stereo sound with time-varying spatial cues.

Training, Development, and Test Sets

For the training, development, and test set, we respectively created 40000, 10000, and 6000 utterances of length 2.4-second using only the simulated BRIRs. For each utterance, we ran-

[3] `http://iosr.uk/software/index.php`

domly sampled a set of BRIRs and two speech samples. The speech signals were rescaled to a random relative SNR between 0 and 5 dB. The moving velocity of each speaker was randomly chosen between 8 and $15°/s$ and the moving direction was randomly chosen between clockwise and counter-clockwise. In addition, to compare the proposed method in different rooms with different velocities, we chose 3 simulated rooms with RT60 0.3 s, 0.5 s, 0.7 s; 3 real rooms with RT60 0.32 s, 0.47 s, 0.68 s; and 3 velocity ranges $5 - 10°/s$, $10 - 15°/s$, $15 - 20°/s$, and generated 1000 utterances on each condition for testing only.

Evaluation Metrics

Similar to the evaluation in Section 2.2, we assessed the models by evaluating both the quality of speech separation and the preservation of spatial cues. SNR was employed as the metric for signal quality. However, ITD and ILD errors are not suitable for utterance-level evaluation of spatial cue preservation because the moving sources are at different locations in an utterance. To address this, we trained a speaker localizer for moving source localization on reverberant clean signals. We set the chunk size as 80 ms, so the localizer predicts the DOA every 80 ms. Since the DOA classes are ordered, even a wrong classification can correspond to a close DOA estimation, e.x., 5° angular error. Therefore, we opted to report the absolute DOA errors as our metric for the accuracy of preserving spatial cues.

Network Architectures

Our MIMO TasNet was built upon the causal configuration of TasNet as detailed in [25]. In the linear encoder and decoder, we employed 64 filters, each with a 4 ms filter length, which corresponds to 64 samples at a 16 kHz sampling rate. The TCN module comprised five repeated stacks, each containing seven 1-D convolutional blocks. This configuration gives the model an effective receptive field of approximately 2.5 seconds. We set the STFT window size to 32 ms and the window shift to 2 ms when calculating cosIPD, sinIPD, and ILD.

Table 2.5: Experimental results of moving source separation in reverberant rooms with various configurations of TasNet. SNR (dB) and DOA error (°).

Method	Context size (s)	SNR (dB)	DOA error (°)
Unprocessed	-	0	-
SIMO TasNet	2.4	5.1	20.9
MIMO TasNet+block-wise	0.1	5.7	16.3
	0.2	6.0	15.4
	0.3	6.2	14.1
MIMO TasNet	2.4	**8.4**	**8.3**
-sinIPD, cosIPD, ILD		7.3	11.0
MISO TasNet	2.4	**9.4**	**6.1**
-mask&sum		8.8	7.3
Reverberant clean	-		0.5

2.3.4 Results and Discussion

Table 2.5 compares different methods for moving source separation. The single-input-multi-output (SIMO) TasNet uses only spectral-temporal information for separation, yielding an average of 5.1 dB SNR improvement. The block-wise adaptation of MIMO TasNet with oracle block tracking achieves better performance than the SIMO TasNet even though it relies on a much shorter context window. This observation suggests the importance of spatial information in source separation. As the duration increases, both SNR and DOA estimation become better. When the MIMO TasNet performs utterance-level separation, it achieved 8.4 dB SNR improvement, outperforming block-wise adaptation by a large margin. The huge improvement confirms the effectiveness of the longer context for moving source separation, as the model could take advantage of longer spectral-temporal and spatial-temporal information. We also notice that the performance of MIMO TasNet drops greatly when the frequency-domain features, i.e., cosIPD, sinIPD, and ILD, are removed. It is likely because the frequency-domain features provide more stable spatial features than those extracted by the parallel encoders in the reverberant environments as STFT uses a longer window size than the linear encoders. In the binaural enhancement stage, the MISO TasNet further improves the SNR and reduces the DOA error. The performance gain from the mask-and-sum mechanism

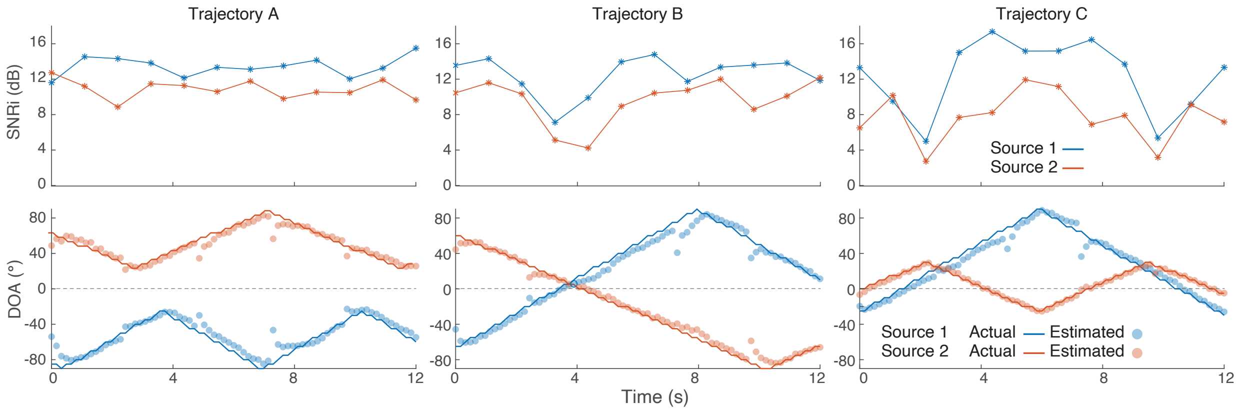


Figure 2.4: SNR improvement (top) and DOA estimation (bottom) over time for two moving speakers on three examples of trajectories in the reverberant room with RT60 0.2s. The result for each trajectory is averaged over 100 instances of speech separation.

shows the effectiveness of combining spatial filtering and spectral filtering to separate sources. Our results show that better signal quality (higher SNR) always leads to better preservation of spatial cues (lower DOA error), which is consistent with the observations in [76, 111] that conducted source separation on nonmoving sources.

Figure 2.4 reports SNR improvement and DOA estimation on three example trajectories. In example A, two speakers move in left and right planes, and the proposed method achieves good performance constantly because two speakers always have distinct spatial information. In B and C trajectories, the SNR improvement becomes less as the two speakers move closer to each other, and improves when they move apart. Interestingly, we notice that the DOA estimation is less affected by speakers' co-location than the SNR. A possible explanation is that the speaker localizer uses the velocity and directional information to compensate for decreased signal quality.

Table 2.6 compares the proposed method on speakers with different moving velocities in different rooms. The model trained in the simulated rooms can generalize to the real rooms well. In both simulated and real rooms, reverberation substantially deteriorates the performance of the model in terms of the signal quality and accuracy of the preserved spatial cues. This degradation is due to the temporal smearing of the mixed signal which combines the components of the same and different speakers over time and makes the mask estimation more challenging. We notice that

Table 2.6: Experimental results of the proposed system with different moving velocity in different rooms. SNR (dB) and DOA error (°) are reported.

Room condition (RT60)		Velocity of motion (°/s)		
		5-10	10-15	15-20
Simulated room	a (0.3 s)	8.8/5.8	8.7/5.6	8.9/6.2
	b (0.5 s)	7.6/7.7	7.6/7.3	7.5/8.3
	c (0.7 s)	6.8/9.2	6.7/9.3	6.7/10.0
Real room	A (0.32 s)	7.5/9.6	7.5/9.9	7.5/9.3
	B (0.47 s)	6.6/15.2	6.5/15.0	6.5/15.2
	C (0.67 s)	6.5/11.9	6.5/12.3	6.3/12.3

velocity has little impact on SNR and, DOA estimation only becomes slightly worse in the highest velocity range. It shows the robustness of the system across various conditions.

2.4 Conclusion

This chapter tackles the challenge of speaker-independent speech separation with an emphasis on low latency. Our first contribution is a real-time single-channel speech separation model, online deep attractor network, which employs an online clustering strategy. We show that the online deep attractor network is competitive with other noncausal models using offline clustering. Then, we propose a real-time binaural speech separation model, MIMO TasNet that simultaneously separate speakers in both the left and right channels. Our experimental results indicate that MIMO TasNet not only achieves superior separation performance but also effectively preserves interaural time difference and interaural level difference in the separated outputs, outperforming existing TasNet variants. Furthermore, we investigated the problem of moving source separation. We upgraded the MIMO TasNet to a two-stage framework that performs utterance-level binaural speech separation and enhancement sequentially. The method fully utilizes long-term spectral and spatial information and implicitly tracks the speakers within the utterance without the need for the external speaker tracking module. Experiments show that the model significantly improves separation performance and more accurately preserves spatial cues compared to the traditional block-wise adaptation method.

Chapter 3: Using Speaker Identifications to Improve Speech Separation

The models introduced in Chapter 2 do not specifically leverage speaker information for speech separation. However, this information can be advantageous in certain situations. For example, in the separation of long recordings, maintaining a consistent allocation of speakers to output channels is a challenge, particularly when speakers move freely. In such contexts, the use of speaker information could facilitate more accurate tracking of speakers. Similarly, in a scenario like a conference room with ten individuals conversing, knowing each participant's unique vocal characteristics could benefit the separation process. This chapter explores the application of speaker information to improve speech separation in these scenarios. Importantly, the methods introduced in this chapter do not rely on pre-recorded samples of the speakers' voices. Instead, they are designed to intelligently extract speaker information directly from the mixed audio.

3.1 Online Binaural Speech Separation of Moving Speakers With a Wavesplit Network

3.1.1 Introduction

We have introduced a deep learning approach that performs utterance-level source separation of moving speakers in Section 2.3. The model utilizes longer spectral and spatial information and implicitly tracks the speakers within the utterance which significantly outperforms block-wise adaptation methods. Although the model shows promising results with sentence-length waveforms, robust tracking moving speakers in long-form speech separation is still challenging, primarily due to the speaker swap problem. The speaker swap problem refers to a scenario in which even though the overlapped sources can be well separated, the ordering of outputs may be inconsistent over time. For example, when separating speakers in a long mixture of A+B, the model outputs [A,B] between time t_1 and t_2 but outputs [B,A] between time t_2 and t_3. This phenomenon is annoying and frequent, and it can occur when speaker energy varies or a period of silence exists amid the mixture. Separating moving speakers is even more prone to the speaker swap problem than separating stationary speakers, especially at times when speakers move closer to each other in space. It is likely because similar spatial information results in speaker tracking ambiguity. To make the speaker order consistent in the sequence, a stitching-based algorithm was proposed, which divides the long-form outputs into several overlapped segments, calculates the similarity between the overlapped regions in adjacent segments, and re-orders the segments [112, 44]. Others designed a tracking network to track the segments [113]. These methods are effective for non-causal systems but not suitable for causal systems which require low latency.

An alternative solution for real-time systems is speaker-informed speech separation. In this method, the model is conditioned on a representation of each speaker which is used to track the speakers over time. The representation is usually a speaker-discriminative embedding such as i-vector [47] or d-vector [48]. There are multiple ways to acquire speaker representations. A general approach is to use a speaker ID neural network to compute speaker embedding from a voice snippet of the target speaker [49, 50, 52, 53]. Speaker representations can also be derived from the sound mixture itself. Wavesplit [114] jointly trains a speaker stack and a separation stack where the speaker stack predicts an embedding per speaker at each time frame, aggregates them across the whole input, and uses the aggregated representation to guide the speaker stack.

In this section, we address the speaker swap problem in the task of binaural speech separation for moving speakers. We propose a new model inspired by the Wavesplit approach. The speaker profile module infers speaker representations from the mixture over time, which are then utilized by the extraction module and localization module to track each speaker reliably. Experimental results show that the proposed method can mitigate the speaker swap problem while achieving comparable performance with u-PIT models with ground truth tracking in both separation quality and preserving the spatial cues in long-form speech separation. Moreover, all the modules in the system are causal and have low latency, making the system suitable for applications in hearable devices.

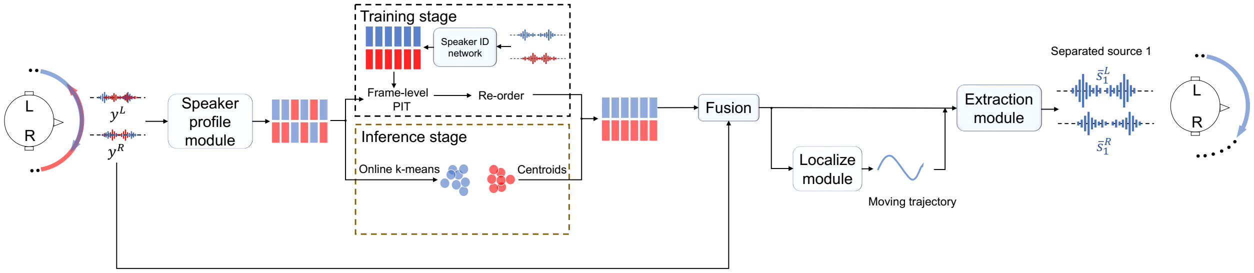


Figure 3.1: Schematic of the proposed system.

3.1.2 Method

Figure 3.1 illustrates the overall flowchart of the proposed framework. The speaker profile module estimates an embedding for each speaker at each time frame from the binaural mixture, but the ordering of the embeddings is not necessarily consistent over time. During training, a pre-trained speaker ID network generates oracle embeddings from individual sources to serve as the training target for the speaker profile module. Frame-level permutation invariant training (PIT) is used to choose the best match and re-order the estimated embedding sequence. In inference, online k-means is performed to cluster embeddings and update the centroids. The re-ordered embedding sequence or the centroid sequence informs the localization and separation modules to jointly localize and separate the corresponding speaker. The interaural cues are preserved in the stereo output.

Speaker Profile Module

Given the binaural mixed signals $\mathbf{Y} \in \mathbb{R}^{2\times L}$, where L is the signal length, the speaker profile module estimates N sequences of speakers vectors, $\mathbf{H} \in \mathbb{R}^{N\times T\times D}$, where N is the number of speakers presented in the mixture, T is the time frames, and D is the vector dimension. In this study, N is fixed to be two. $\mathbf{h}(n,t) \in \mathbb{R}^D$ denotes the n-th vector of $\mathbf{H}$ at the time frame t. It is noted that there is no predetermined ordering of the speaker embeddings at each time frame and the ordering across frames is not necessarily consistent. For example, $\mathbf{h}(1,t_1)$ and $\mathbf{h}(2,t_3)$ can represent the speaker A while $\mathbf{h}(2,t_1)$ and $\mathbf{h}(1,t_3)$ represent the speaker B as shown in Figure 3.1. The embeddings are speaker-discriminative so that 1) they can be clustered into individual speaker groups; 2) each group of embeddings can guide the separation of the corresponding speaker from the mixture.

To facilitate training, we made some modifications to Waveplit. Instead of maintaining a speaker embedding table, we trained a speaker ID network (SNet) to extract the oracle speaker embeddings from the individual sources. SNet follows the design in Zhou et al. [115] and is trained to predict the M different speaker identities using the cross-entropy loss. The SNet takes the source i, $\mathbf{s}_i \in \mathbb{R}^L$, as input and outputs an embedding sequence $\mathbf{E_i} \in \mathbb{R}^{T\times D}$. $\mathbf{E} = [\mathbf{E}_1, ..., \mathbf{E}_N] \in \mathbb{R}^{N\times T\times D}$ is the oracle speaker embeddings. Different from $\mathbf{H}$, the ordering of speakers in $\mathbf{E}$ is consistent in the sequence. To force the embeddings to have small intra-speaker and large inter-speaker distances, we randomly sample time frames $\{p, q\}$ and add a triplet loss,

$$\begin{aligned}\mathcal{L}_{\text{triplet}} = \sum_{i,j}\sum_{p,q} \max\{&|\mathbf{e}(i,p) - \mathbf{e}(i,q)| - \\ &|\mathbf{e}(i,p) - \mathbf{e}(j,p)| + \mathrm{m}, 0\},\end{aligned} \tag{3.1}$$

where $\mathbf{e}(i, p)$ is the vector of speaker i at the time step p and m is the margin. Then, the frame-level PIT loss is used to train the speaker profile module,

$$\mathcal{L}_{\text{PIT}}(\mathbf{H}, \mathbf{E}) = \sum_{t=1}^{T} \min_{\pi\in P} \sum_{i=1}^{N} |\mathbf{h}(i,t) - \mathbf{e}(\pi(i), t)|, \tag{3.2}$$

where P is the set of all N! permutations. The best match between the oracle embeddings and estimated embeddings at each frame can be used to re-order the estimated speaker embeddings in an order consistent with the oracle ones. The re-ordered embeddings $\hat{\mathbf{H}}(i) \in \mathbb{R}^{T\times D}, i = 1...N$ is the i-th speakers' profile used to guide the speech separation.

In inference, online k-means is performed on $\mathbf{H}$, which keeps updating the cluster centroids over time. The sequence of the centroids $\mathbf{C}(i) \in \mathbb{R}^{T\times D}$ is the i-th speaker' profile.

Speaker Localization Module and Extraction Module

We train a multi-input-multi-output (MIMO) TasNet to separate the targeted speaker. Following the design in Section 2.3, we use 1-D convolution layers to extract the time-domain features from the waveform, and then concatenate the time-domain features and frequency-domain features, i.e., interaural phase difference (IPD) and interaural level difference (ILD) as input features which contain spectro-temporal and spatial-temporal information. The difference is that the model is conditioned by a speaker profile. Previous research has shown that feature-wise linear modulation (FiLM) [116] is an effective conditioning method for neural networks, so we use the same method here. Specially, FiLM learns two linear functions $f_l \in \mathbb{R}^{D\times D}$ and $g_l \in \mathbb{R}^{D\times D}$ at layer l which project the speaker profile $\hat{\mathbf{H}}(i)$ to $\gamma_l(i) \in \mathbb{R}^{T\times D}$ and $\beta_l(i) \in \mathbb{R}^{T\times D}$, respectively,

$$\gamma_l(i) = \hat{\mathbf{H}}(i) \cdot f_l, \quad \beta_l(i) = \hat{\mathbf{H}}(i) \cdot g_l, \tag{3.3}$$

where $\gamma_l(i)$ and $\beta_l(i)$ modulate the activation x_l at layer i,

$$\text{FiLM}(x_l|\gamma_l(i), \beta_l(i)) = \gamma_l(i) \times x_l + \beta_l(i). \tag{3.4}$$

We add one FiLM before each convolutional block in TasNet.

The system jointly localizes and separates the target speaker. The localization module performs the classification of the direction of arrival (DOA) at each time frame. The DOA angles are discretized into K classes. The localization module estimates a classification matrix $\mathbf{V}(i) \in \mathbb{R}^{T\times K}$ for the speaker i. To train the localization module, we compute the cross-entropy loss between $\mathbf{V}(i)$ and the target DOA labels. We split $\mathbf{V}(i)$ into B small chunks with each chunk containing Q time frames, $T = B \times Q$, and count the frequency of each DOA class in each chunk, and consider the most frequent class as the estimated DOA for that chunk. The explicitly estimated trajectories enable us to determine the moving source we are interested in and to modify the acoustic scene accordingly, for example, by amplifying or attenuating individual sources.

We concatenate $\mathbf{V}(i)$ with other fusion features to extract the target speaker $\hat{\mathbf{S}}_i = [\hat{\mathbf{s}}^L, \hat{\mathbf{s}}^R]$, where $\hat{\mathbf{s}}_i^L$ and $\hat{\mathbf{s}}_i^R$ are the estimated left- and right-channel signals of the source i. Since the target speaker is determined by $\hat{\mathbf{H}}(i)$, there is no permutation problem. The reconstruction loss $\mathcal{L}_{\text{extraction}}$

is:

$$\mathcal{L}_{\text{extraction}} = \text{SNR}(\mathbf{s}_i^L, \hat{\mathbf{s}}_i^L) + \text{SNR}(\mathbf{s}_i^R, \hat{\mathbf{s}}_i^R), \tag{3.5}$$

$$\text{SNR}(\mathbf{x}, \hat{\mathbf{x}}) = 10 \log_{10} \left(\frac{||\mathbf{x}||_2^2}{||\hat{\mathbf{x}} - \mathbf{x}||_2^2} \right). \tag{3.6}$$

The speaker profile module, localization module, and extraction module were trained separately and then trained jointly.

3.1.3 Experimental Settings

Dataset

The binaural room impulse responses, speech data, and moving source simulation method used here are consistent with those described in Section 2.3.3.

Evaluation Metrics

In addition to assessing separation quality and spatial cue preservation, we also evaluated the model's robustness against the speaker swap problem. Following [41], we trained a speaker localizer to examine the locational information encoded in stereo outputs. The localizer predicts the DOA every 80 ms. We calculated absolute DOA errors as the metric for the accuracy of preserving spatial cues. It is noted the localization model for evaluation here is different from the one in Section 3.1.2 as the latter one aims to decode the trajectory of the target speaker to facilitate the separation. When evaluating long-form speech separation, we divided the separation outputs into N segments and checked if the order of outputs in adjacent segments was consistent. The number of speaker swaps within the separation outputs was counted and used as a metric to gauge the model's robustness against the speaker swap issue. Figure 3.2 shows three examples. The duration of long recordings is 24 seconds, and N is 10.

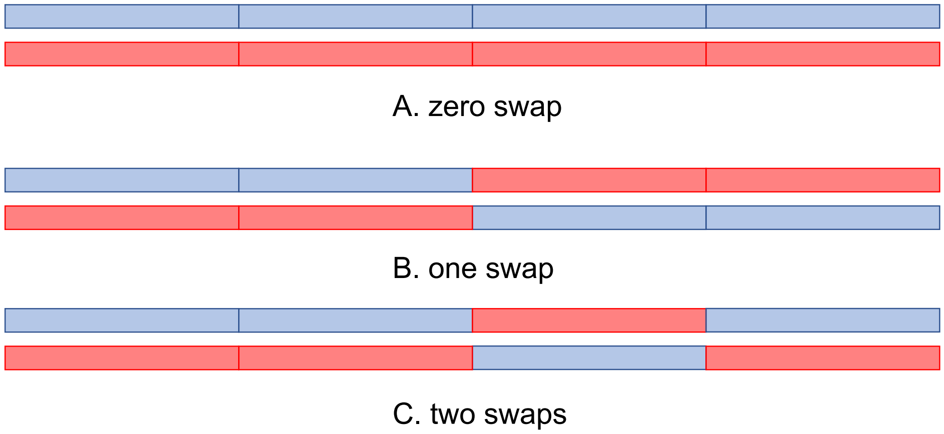


Figure 3.2: Examples of speaker swap in separation outputs. Blue bar denotes the speaker 1 and red bar denotes the speaker 2.

Network Architectures

All the modules, except for the localization module, were built upon the causal configuration of TasNet [25]. We used 64 filters in the linear encoder and decoder with a 4-ms filter length (i.e., 64 samples at 16 k Hz). We used 5 repeated stacks for the speaker profile module and speaker ID network, 2 repeated stacks for the fusion module, and 3 repeated stack for the extraction module with each stack having seven 1-D convolutional blocks. The localization module is a two-layer uni-directional LSTM.

Models for Comparison

We used several monaural and binaural separation models for comparison. The single-input-multi-output (SIMO) TasNet is a monaural model that separates the mixture in left- and right-channels independently. Block-wise MIMO TasNet and uPIT-MIMO TasNet (in Section 2.3) are binaural baselines. Block-wise MIMO TasNet separates speech in each short block independently and concatenates the block outputs using oracle tracking information. SPK-MIMO TasNet is the proposed method in this chapter, and oracle SPK-MIMO TasNet means the model is conditioned on the oracle speaker profiles from the pre-trained speaker ID network. We add the same post binaural speech enhancement (in Section 2.3) after each binaural separation model for comparison, denoted

Table 3.1: Experimental results of moving source separation on 2.4-second recordings.

Method	Context size (s)	SNR (dB)	DOA error (°)
Unprocessed	-	0	-
SIMO TasNet	2.4	5.1	20.9
Block-wise MIMO TasNet	0.1	5.7	16.3
	0.2	6.0	15.4
	0.3	6.2	14.1
uPIT-MIMO TasNet	2.4	8.4	8.3
SPK-MIMO TasNet		8.3	8.2
Oracle SPK-MIMO TasNet		8.9	7.4
uPIT-MIMO TasNet + Enh	2.4	9.4	6.1
SPK-MIMO TasNet + Enh		9.4	6.0
Oracle SPK-MIMO TasNet + Enh		9.6	5.8
Reverberant clean	-		0.5

as “+enh" in Table 3.1. When evaluating models on long recordings, we divided the separation outputs of uPIT-MIMO TasNe into segments as shown in Figure 3.1 and re-ordered the segments using the ground truth signals, which is referred to as uPIT-MIMO TasNet w/ tracking in Table 3.2 and Table 3.3.

3.1.4 Results and Discussion

Table 3.1 compares different methods for moving source separation on 2.4 s recordings. uPIT-MIMO TasNet outperforms both the SIMO TasNet and block-wise adaptation methods by a large margin because uPIT-MIMO TasNet takes advantage of longer spectral-temporal and spatial-temporal information for moving speaker separation. Oracle SPK-MIMO TasNet, conditioned on the oracle speaker profile, achieves 0.5 dB SNR gain over uPIT-MIMO TasNet, proving the effectiveness of speaker-informed speech separation in moving speaker cases. The performance drops slightly when we use the speaker profiles inferred from the mixture. It is different from the results in Zeghidour et al. [114] where using inferred speaker representation greatly outperforms uPIT-based models. One explanation is that Wavesplit employs various forms of regularization to improve the generalization capability during training. The other is that inferring long-form speaker represen-

Table 3.2: Experimental results on 24-second long recordings.

Method	# of swaps	SNR (dB)	DOA error (°)
uPIT-MIMO TasNet w/o tracking	3.4	1.2	35.3
uPIT-MIMO TasNet w/ tracking	3.4	7.6	10.0
SPK-MIMO TasNet	0.6	7.7	9.3
oracle SPK-MIMO TasNet	0.4	8.2	8.1

tations that are only related to voice characteristics from moving speakers is more challenging, especially in reverberant environments. Moreover, our system is in causal configuration, and the speaker profiles are updated over time, so the speaker profiles become stable after several time frames. In real applications, we can select the source of interest based on the speaker representation, decoded moving trajectory, and the separated waveforms and employ a post binaural speech enhancement module to enhance the result. We notice that a post binaural speech enhancement module lets both uPIT-MIMO TasNet and SPK-MIMO TasNet improve the SNR and reduce the DOA error.

Table 3.2 compares methods on 24 s recordings. We see that uPIT-MIMO TasNet has multiple times of speaker swaps in long-form speech separation, which severely affects the overall SNR. The speakers are separated but not consistently placed in certain output channels. With the ground truth tracking, the overall SNR is improved from 1.2 dB to 7.5 dB and the overall DOA error is reduced from 35.3°to 10.0°. We see that SPK-MIMO TasNet and oracle SPK-MIMO TasNet, which are conditioned on a speaker representation, are less prone to speaker swap and tend to assign each speaker to a certain output channel. The model conditioned on the inferred speaker profiles is slightly worse than that conditioned on the oracle speaker profile but is slightly better than the u-PIT based model with tracking. This shows the proposed method is more robust with respect to a targeted moving signal.

We also compared the models on 24 s recordings simulated using real room impulse responses with different reverberant time as shown in Table 3.3. All the models trained in the simulated rooms can generalize to the real rooms well. Stronger room reverberation deteriorates the performance of

Table 3.3: Experimental results on 24-second long recordings in different rooms. The number of swaps / SNR (dB) / DOA error (°) are reported.

Model	Rooms		
	A (0.32 s)	B (0.47 s)	C (0.67 s)
uPIT-MIMO TasNet w/o tracking	3.6/1.3/37.2	3.5/1.2/36.4	3.6/1.0/38.0
uPIT-MIMO TasNet w/ tracking	3.6/7.0/11.6	3.5/6.1/15.1	3.6/6.0/16.3
SPK-MIMO TasNet	0.6/7.2/10.0	0.8/6.3/13.4	0.6/6.2/13.2
oracle SPK-MIMO TasNet	0.3/7.6/9.6	0.4/6.5/12.1	0.4/6.4/12.9

the model in terms of the signal quality and accuracy of the preserved spatial cues. However, room reverberation has little impact on the number of speaker swaps.

3.2 Continuous Speech Separation Using Speaker Inventory for Long Recording

3.2.1 Introduction

In Section 3.1, we have demonstrated that speaker embeddings can enhance the model's ability to robustly track speakers. Besides this study, leveraging speaker information has received increasing attention [49, 50, 52, 53, 117, 118, 119]. We can categorize them into two main categories. The first category is informed speech extraction, which exploits an additional voice snippet of the target speaker to distinguish his/her speech from the mixture. SpeakerBeam [49, 50] derives a speaker embedding from an utterance of the target speaker by using a sequence summary network [51] and uses the embedding to guide an extraction network to extract the speaker of interest. VoiceFilter [52] concatenates spectral features of the mixture with the d-vector [48] of a voice snippet to extract the target speaker. Xiao et al. [53] used an attention mechanism to generate context-dependent biases for target speech extraction. Informed speech extraction solves the permutation problem and does not need to predetermine the number of outputs. However, it has two limitations. Firstly, the computation cost is proportional to the number of speakers to be extracted, so in a multi-speaker conversation, the system needs to run multiple times to extract each speaker one by one. Most importantly, the extraction usually fails when the target speaker's biased information is not strong enough [50].

The second category is speech separation using speaker inventory (SSUSI) [119]. The method employs a pool of additional enrollment utterances from a list of candidate speakers, from which profiles of relevant speakers involved in the mixture are first selected. Then the method fuses the selected profiles and the mixed speech to separate all speakers simultaneously. As multiple profiles are provided during separation, more substantial speaker discrimination can be expected, which yields better speech separation. The method can also employ permutation invariant training (PIT) [22] to compensate for weak biased information and wrong selection.

Though with prior promising results, methods in both categories assume additional speaker information is available ahead of extraction or separation, which may be impractical in real scenarios. Wavesplit approach, as we adopt in Section 3.1, infers speaker embeddings for each source at each time step from the mixture, aggregates them across the whole input, and uses aggregated representation to guide speaker separation. However, most methods mentioned above prove their successes on fully overlapped speech. The practicality of these methods is unclear as the overlap in real conversation usually possess very different characteristics [120, 121, 122, 123].

In this section, we address these problems on the continuous speech separation (CSS) task [124, 112]. CSS focuses on separating long recordings where the overall overlap ratio is low and the speaker activations are sparse. A large amount of non-overlapped regions in the recording enables the derivation of robust features for the participants. We adopt the SSUSI in the CSS task and propose continuous SSUSI (CSSUSI), which constructs the speaker inventory from the mixed signal itself, instead of external speaker enrollments, by using speaker clustering methods. CSSUSI informs the separation network with relevant speaker profiles dynamically selected from the inventory to facilitate source separation at local regions. The outputs from local regions are then concatenated such that the output audio streams are continuous speech that do not contain any overlap. We created a more realistic dataset that simulates natural multi-talker conversations in conference rooms to test CSSUSI on the CSS task. Experimental results show that CSSUSI can build a speaker inventory from the long speech mixture using the clustering-based method and take advantage of the global information to improve separation performance significantly.

3.2.2 Method

SSUSI Using Pre-Enrolled Utterance

We first overview the original SSUSI system [119], which requires pre-enrolled speaker signals. A SSUSI system contains three modules: a speaker identification module, a speaker profile selection module, and a biased speech separation module. The speaker identification module is responsible for extracting embeddings from both the speaker enrollments and input mixture. Embeddings of speaker enrollments are used for speaker inventory construction. The speaker profile selection module selects from the inventory the best-matched speaker profiles with the mixture embeddings. The selected profiles are then fed into the biased separation module to separate speakers in the mixture.

Since each speech segment is short (4 s in this study) and typically contains at most two speakers, we focus on two-speaker separation for each segment, and the model always generates two outputs. Moreover, we make several modifications to the original SSUSI architecture [119] for better performance.

The speaker identification module constructs the speaker inventory first. The inventory is a pool of K-dimensional speaker embeddings $\left\{\mathbf{e}^j\right\}_{j=1}^{M}, \mathbf{e}^j \in \mathbb{R}^K$, which are extracted from a collection of time-domain enrollment speech $\left\{\mathbf{a}^j\right\}_{j=1}^{M}, \mathbf{a}^j \in \mathbb{R}^{L_{a_j}}$, where L_{a_j} is the temporal dimension of speech signal $\mathbf{a}^j$. M is typically larger than the maximum number of speakers in the mixture to be separated. We also assume that each speaker only has one enrollment sentence. The embedding $\mathbf{E}^j \in \mathbb{R}^{T_j \times K}$ is extracted from $\mathbf{a}^j$ by a speaker identification network (SNet), where T_j is the temporal dimension of the embedding sequence. Here we simply use mean-pooling across the T_j frames of $\mathbf{E}^j$ to obtain the single vector $\mathbf{e}^j \in \mathbb{R}^K$. The mixture embeddings $\mathbf{E}^y \in \mathbb{R}^{T_y \times K}$ are directly extracted from the input mixture $\mathbf{y} \in \mathbb{R}^T$ by SNet , where T and T_y are the temporal dimension of the input mixture and the mixture embeddings, respectively.

The speaker profile selection module selects the relevant speaker profiles from the inventory that are best matched with the mixture embeddings $\mathbf{E}^y$. The selection is performed by calculating the similarity between the mixture embeddings and items in the inventory, and two items with the highest similarity are selected. The similarity are calculated by applying the Softmax function on the dot-product between the mixture and inventory embeddings:

$$\begin{aligned}\mathbf{d}_s^{y,j} &= \mathbf{e}_s^{y} \cdot \mathbf{e}^{j} \\ \mathbf{w}_s^{y,j} &= \frac{\exp(\mathbf{d}_{\mathrm{s}}^{\mathrm{y,j}})}{\sum_{p=1}^{M} \exp(\mathbf{d}_{\mathrm{s}}^{\mathrm{y,p}})}\end{aligned} \tag{3.7}$$

where $\mathbf{e}_s^y$ denotes $\mathbf{E}^y$ at temporal index s. We then calculate the average score $\mathbf{w}^{y,j}$ across the T_y frames:

$$\mathbf{w}^{y,j} = \frac{\sum_{s=1}^{T_y} \mathbf{w}_s^{y,j}}{T_y} \tag{3.8}$$

Two inventory items $\mathbf{e}^{p_1}$ and $\mathbf{e}^{p_2}$ are then selected according to the two highest scores in $\left\{\mathbf{w}^{y,j}\right\}_{j=1}^{M}$.

The biased speech separation module is then adapted to the speech characteristics of the speakers selected from the inventory. This module contains three layers, a feature extraction layer, a profile adaptation layer, and a separation layer. Both feature extraction and separation layers are 2-layer BLSTM in this study. Previous research [49] has shown that a multiplicative adaptation layer, i.e., multiplying the speaker embedding with the output of one of the middle layers of the network, is a simple yet effective way to realize adaptation, so we use the same method here. Given the two selected speaker profiles $\mathbf{e}^{p_1}$ and $\mathbf{e}^{p_2}$, two target-biased adaptation features are calculated by frame-level element-wise multiplication between the profiles and the output of the feature extraction layer:

$$\mathbf{a}_l^{p_1} = \mathbf{b}_l \odot \mathbf{e}^{p_1} \tag{3.9}$$

$$\mathbf{a}_l^{p_2} = \mathbf{b}_l \odot \mathbf{e}^{p_2} \tag{3.10}$$

where $\mathbf{b}_l \in \mathbb{R}^K$ denotes the output of the feature layer, l denotes the frame index, and $\odot$ denotes

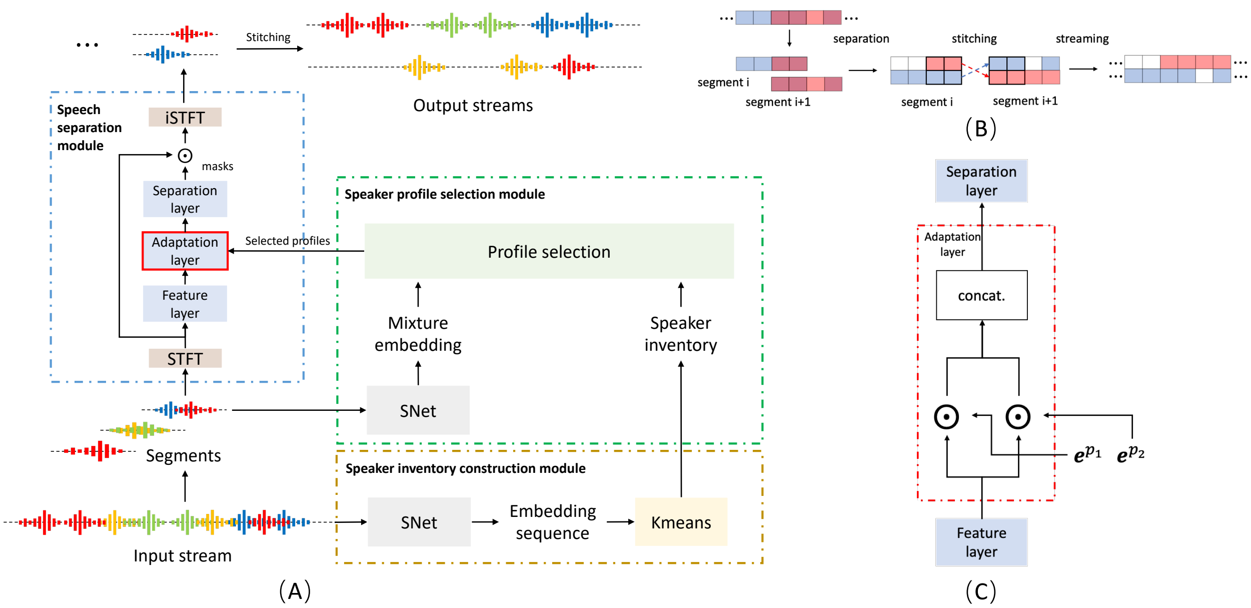


Figure 3.3: (A) The architecture of the proposed continuous speech separation using speaker inventory. The Speaker inventory construction module forms the speaker inventory from the long mixture by using Kmeans clustering; the long mixture is split into small segments, and the speaker profile selection module selects two relevant profiles from the inventory for each segment; the speech separation module fuses the selected speaker profiles into the system for source separation. (B) Stitching procedure of adjacent segment outputs in a long recording. (C) Multiplicative adaptation of the selected profiles $\mathbf{e}^{p_1}$ and $\mathbf{e}^{p_2}$.

the element-wise multiplication. The two target-biased features are then concatenated:

$$\mathbf{A} = \text{concat}([\mathbf{A}^{\text{p}_1}, \mathbf{A}^{\text{p}_2}]) \tag{3.11}$$

where $\mathbf{A}^{p_1} = [\mathbf{a}_1^{p_1}, \ldots, \mathbf{a}_L^{p_1}] \in \mathbb{R}^{L\times K}$, $\mathbf{A}^{p_2} = [\mathbf{a}_1^{p_2}, \ldots, \mathbf{a}_L^{p_2}] \in \mathbb{R}^{L\times K}$, and $\mathbf{A} \in \mathbb{R}^{L\times 2K}$. The separation layer takes $\mathbf{A}$ as the input and estimates two time-frequency (T-F) masks $\mathbf{M}^1, \mathbf{M}^2 \in \mathbb{R}^{L\times F}$.

Continuous SSUSI Using Self-Informed Mechanism for Inventory Construction

SSUSI assumes that pre-recorded utterances of all speakers are available for the speaker inventory construction. However, such an assumption may not be realistic, especially for unseen speakers or meeting scenarios where the collection of pre-recorded speech from the participants is not feasible.

Continuous speech separation (CSS) aims at estimating individual target signals from a continuous speech with a long duration. The continuous speech contains both overlapped and non-overlap speech, and the overlap ratio is relatively low. Therefore, the single-speaker regions can be exploited to derive robust acoustic characteristics of participating speakers without the need for external utterances, which makes the self-informed speaker inventory construction possible. This section introduces how we adopt SSUSI in the CSS task and eliminate the need for pre-recorded speech by using a clustering method.

Figure 3.3 (A) shows the overall flowchart of the continuous SSUSI (CSSUSI) framework. The main difference between CSSUSI and the original SSUSI is in the construction of the speaker inventory. Original SSUSI applies the speaker identification module on extra enrollment utterances, whereas CSSUSI first splits the mixture recording $\mathbf{y}$ into B small chunks, and directly extracts the mixture embeddings $\{\mathbf{e}_b^y\}_{b=1}^B$, where $\mathbf{e}_b^y \in \mathbb{R}^K$ denotes the embedding vector in chunk b. Then, CSSUSI applies Kmeans clustering on $\{\mathbf{e}_b^y\}_{b=1}^B$ to form M clusters, and the cluster centroids form the speaker inventory. In Section 3.2.4 we will show that the separation performance is insensitive to the choice of M as long as M is no smaller than the actual number of active speakers in the recording.

CSUSSI uniformly segments the mixture recording and exploits the inventory to facilitate source separation in each segment. Except for the self-informed speaker inventory, CSSUSI uses the same speaker profile selection and biased speech separation methods as introduced in SSUSI above, respectively. To stitch the outputs from each segment into output streams where each stream only contains non-overlapped speakers, the similarity between the overlapped regions in adjacent segments determines which pair to be stitched. Figure 3.3 (B) shows the stitching procedure of adjacent segments.

3.2.3 Experimental Settings

Dataset

In our training set, we randomly generated 3000 rooms. The length and width of the rooms are randomly sampled between 5 and 12 meters, and the height is randomly sampled between 2.5 and 4.5 meters. A microphone is randomly placed in the room, and its location is within 2 meters of the room center. The height of the microphone is randomly sampled between 0.4 and 1.2 meters. We randomly sample 10 speakers from the LibriSpeech corpus [125] for each room. All the speakers are at least 0.5 meters away from the room walls and the height of the speakers are between 1 and 2 meters. The reverberation time is uniformly sampled between 0.1 and 0.5 seconds. We randomly chose 2 speakers as relevant speakers and arrange them according to one of the four following patterns:

1. *Inclusive*: one speaker talks a short period while the other one is talking.
2. *Sequential*: one talks after the other one finishes talking.
3. *Fully-overlapped*: two speakers always talk simultaneously.
4. *Partially-overlapped*: two speakers talk together only in a certain period.

We selected four patterns with respective frequencies of 10%, 20%, 35%, and 35%. The minimal length of the overlapped periods in inclusive and partially-overlapped patterns is set to 1 second. The maximal length of the silent periods between the two speakers in the sequential pattern is 0.5 second. Moreover, to generate single-speaker utterances, there is a 0.1 probability that one of the speakers is muted in each pattern. We used the remaining 8 speakers as the irrelevant speakers that will not appear in the mixture. Each of the room configurations is used for 8 times. The mixture length is 4 seconds. So, the total training time is $3000 \times 8 \times 4s = 26.7$ hours. For both the relevant and irrelevant speakers, a 10-second utterance is sampled to form the speaker inventory. All speech signals are single-channel and sampled at 16 kHz. Gaussian noise with SNR randomly chosen between 0 and 20 dB is added into the mixture.

Table 3.4: SNR (dB) on eight-speaker long recordings (segment-wise evaluation). The performance on different overlap ratios is reported.

Method	Speaker enrollment	Overlap ratio in %					
		0	0-25	25-50	50-75	75-100	Average
Unprocessed	-	8.6	-9.7	-1.2	-0.9	-0.7	-0.1
BLSTM	-	15.5	8.0	8.6	7.5	6.9	10.6
SSUSI	Two wrong profiles	15.2	7.1	8.4	7.8	7.1	10.3
	One correct and one wrong profiles	15.4	7.8	9.0	8.2	7.6	10.7
	Two correct profiles	15.9	9.5	10.6	9.4	8.7	11.9
	Selected profiles	15.7	8.8	10.0	9.0	8.3	11.5

In our testing set, we set three configurations: 60-second mixture containing 2 speakers, 150-second mixture containing 5 speakers, and 240-second mixture containing 8 speakers. We generated 300 recordings for each configuration. The overall overlap ratio of each recording is 30% complying with natural conversion [126].

Implementation Details

All the models contained 4 bidirectional LSTM (BLSTM) layers with 600 hidden units in each direction. The speaker identification module, which adopts a previous design [115], was pretrained on the VoxCeleb2 dataset [127] and achieved 2.04% equal error rate on the VoxCeleb1 test set [128]. The module extracts 128-dimensional speaker embeddings for every 1.2-second (30-frame) segment. We used SNR as the objective function [40] for separation modules and Adam [85] as the optimizer with initial learning rate of 0.001. The learning rate was decayed by 0.98 for every two epochs.

3.2.4 Results and Discussion

Table 3.4 compares different DNN models on 4-second segments of eight-speaker recordings. The inventory contains eight speakers' profiles that are derived from eight external utterances. SSUSI achieves leading performance on all levels of overlap ratios when two correct speaker profiles are used; however, the performance of SSUSI drops greatly with two wrong speaker profiles

Table 3.5: SNR (dB) on long recordings with different configurations (segment-wise evaluation).

# Speaker	Method	External utterances	Clusters	Avg.
2 speakers	Unprocessed	-	-	1.6
	BLSTM	-	-	11.2
	SSUSI	2	No	12.2
	CSSUSI	No	2	12.1
			3	11.9
			4	11.9
5 speakers	Unprocessed	-	-	0
	BLSTM	-	-	10.6
	SSUSI	5	No	11.5
	CSSUSI	No	3	10.9
			5	11.3
			8	11.2
			10	11.2
8 speakers	Unprocessed	-	-	-0.1
	BLSTM	-	-	10.6
	SSUSI	8	No	11.5
	CSSUSI	No	5	11.0
			8	11.3
			12	11.3
			16	11.2

randomly chosen from the 8 irrelevant speakers, which indicates that performance gain obtained by SSUSI mainly comes from leveraging the target speaker information. We also notice that the performance of SSUSI with two wrong profiles is only slightly worse than the baseline BLSTM, and when only one correct speaker profile is enrolled, SSUSI can still outperform the baseline model, which proves that PIT can compensate for wrong selection and the separation module is robust to adaptation features. When the speaker profiles are selected by the profile selection module, the SSUSI model performs slightly better on the non-overlapped mixtures (overlap ratio is 0) but much better on the overlapped mixtures at all overlap ratios. This confirms the effectiveness of the SSUSI framework on improving separation performance across various settings, which is consistent with previous experimental results on Librispeech although the model architectures are different [119].

Table 3.5 compares CSSUSI with different clusters on recordings with different number of

Table 3.6: Utterance-level evaluation. SI-SDR(dB) is reported.

Method	Need external utterances?	2 spk	5 spk	8spk
Unprocessed	-	6.0	4.5	4.3
BLSTM	No	11.7	10.8	10.6
SSUSI	Yes	13.2	12.0	11.7
CSSUSI	No	13.1	11.9	11.7

speakers. Since the number of participating speakers in a meeting may be unknown, we intend to do over-clustering, i.e., setting the number of clusters greater than the number of speakers in a meeting. Table 3.5 compares CSSUSI with different clustering settings. The performance of CSSUSI is almost identical once the number of clusters is not fewer than the number of speakers. Over-clustering has very little impact on the performance as it ensures each speaker possesses at least one cluster center. Some extra clusters may represent acoustic characteristics of overlapped regions, which will be regarded as irrelevant profiles during profile selection. We see that CSSUSI outperforms the baseline model BLSTM on all configurations. As we conclude from Table 3.4, the performance gain is achieved via leveraging relevant speakers' information. So the performance gain from CSUSSI suggests the successful construction of the speaker inventory from the mixture itself and effective utilization of speaker information. Furthermore, we compare CSSUSI with SSUSI that derives speaker profiles from external utterances that contain only a single speaker in each utterance. CSSUSI sacrifices very little performance but does not require external utterances, which shows CSSUSI is a better model than SSUSI for long recording speech separation.

Table 3.6 compares utterance-wise separation performance. After segments are stitched, each complete utterance is extracted from the output streams by using ground-truth segmentation information, i.e., onset and offset of each utterance. We find that CSSUSI surpasses the baseline in all configurations by a large margin, which further proves that the speaker embeddings can be derived from the raw mixture and enhance speech separation in the long recordings.

3.3 Conclusion

This chapter investigates using speaker information to improve speech separation in certain situations. First, we addressed the problem of binaural speech separation of moving speakers while preserving interaural cues for long recordings. We propose a Wavesplit-based model that estimates speaker embeddings per speaker and per frame from the mixture, performs online clustering to aggregate the embeddings into individual speaker profiles, and conditions on each speaker profile to localize and separate the speaker faithfully. Objective evaluations demonstrate that the proposed model mitigates the swap problem while achieving on par performance with uPIT-based models with ground truth tracking.

Second, we present continuous speech separation using speaker inventory for long audio recordings of multi-talker meetings. Recognizing that meeting audio recordings generally contain a large amount of non-overlapped regions, we propose CSSUSI that can construct a speaker inventory from the long recordings and then utilize the inventory to improve speech separation. CSSUSI overcomes the limitation of the original SSUSI that requires external enrollments. Experiments on a simulated noisy reverberant dataset show that CSSUSI significantly outperforms the baseline models across various conditions.

The application of speaker information, however, is not confined to these two scenarios. Future research will investigate its broader application across a more diverse range of situations.

Chapter 4: Developing Separation Models Using Real Recordings

In this chapter, we introduce using unsupervised learning methods to improve speech separation performance on real recordings. In the first section, we recorded audio data with reverberation and moderate environmental noise using a pair of microphone arrays placed around each of the two ears and then mixed sound recordings to simulate adverse acoustic scenes. Then, we trained a multi-channel speech denoising network (MCSDN) on the mixture of recordings. To improve the training, we employ an unsupervised method, complex angular central Gaussian mixture model (cACGMM), to acquire cleaner speech from noisy recordings to serve as the learning target. We propose a MCSDN-Beamforming-MCSDN framework in the inference stage. The results of the subjective evaluation show that the cACGMM improves the training data, resulting in better noise reduction and user preference, and the entire system improves the intelligibility and listening experience in noisy situations.

In the second section, we extend the recently-proposed mixture invariant training (MixIT) algorithm to perform unsupervised learning in the multi-channel setting. We use MixIT to train a model on far-field microphone array recordings of overlapping reverberant and noisy speech from the AMI Corpus. The models are trained on both supervised and unsupervised training data, and are tested on real AMI recordings containing overlapping speech. To objectively evaluate our models, we also use a synthetic multi-channel AMI test set. Holding network architectures constant, we find that a fine-tuned semi-supervised model yields the largest improvement to SI-SNR and to human listening ratings across synthetic and real datasets, outperforming supervised models trained on well-matched synthetic data. Our results demonstrate that unsupervised learning through MixIT enables model adaptation on both single- and multi-channel real-world speech recordings.

4.1 Refine Supervised Training Data Using Gaussian Mixture Model

4.1.1 Introduction

Removing noise from speech signals is a good way to improve a user's experiences in noisy environments. New hardware allows for multiple microphones near the ear and the processing power to learn from these signals, which can deliver a better auditory experience. This work describes a system for denoising speech signals captured by a pair of microphone arrays near the ears under noisy conditions. We capitalize on deep neural network (DNN) architectures for speech enhancement, along with multi-channel beamforming.

DNN training requires a large quantity of realistic training data. For speech enhancement, the labeled data is a pair of noisy and clean speech signals. One can create arbitrary amounts of noisy data by adding reverberation and noise; however there is still a mismatch between simulated noisy mixtures and real-world audio due to the complex acoustics of real environments. This could degrade the DNN's performance when it is applied to real-world data.

In this work we instead measured a large quantity of speech and noise signals in a real room, and create mixtures from these recordings. This is good for realism but also includes room tone. An important part of this work is a method for preprocessing the recorded sound to remove this background noise so that it can be used as ground truth. We demonstrate improvements in noise reduction and listening preference due to the preprocessing.

Related Work

Speech enhancement has been actively studied for decades [129, 130, 131, 37]. In recent years, deep neural networks have greatly advanced speech enhancement using both supervised [132, 133, 134] and unsupervised methods [56, 57, 58, 59, 135]. Supervised methods have achieved overwhelming performance, but they require access to ground-truth signals, and thus they can degrade on real recordings which are mismatched with the simulated data used for training. Unsupervised methods overcome these problems by requiring only the noisy speech signals.

A general category of unsupervised approaches utilizes spatial information to cluster sound sources in space [56, 57, 58, 59]. The posterior cluster labels can be used as masks to isolate the target speech. An approach using the complex angular-central Gaussian mixture model (cACGMM) [57] clusters the signals, and the resulting labels are used as pseudo-target to train a deep clustering model [20].

This chapter employs cACGMM to extract cleaner speech signals from the recordings to serve as the training target. Our motivation is that we can easily collect moderately noisy recordings without access to ground-truth signals in real scenarios, which can be well processed by the unsupervised clustering methods. Then, we mix several recordings into a much noisier mixture and take advantage of supervised learning to predict the clean speech signals from the mixture. The difference compared to prior work is that we do not apply the clustering model to the noisy mixture directly, because the clustering-based methods perform poorly in challenging conditions where spatial features are smeared by room reverberance and strong background noise, especially diffuse noise with no distinct directional features.

Using a DNN to predict the masks that estimate the spatial covariance, which steers the beamformer toward the target signal, is a popular method to combine DNNs and conventional beamforming methods [136, 137, 138]. The linear beamformers effectively keep speech free of nonlinear distortion, which is essential for good perceptual quality of speech in communication. However, the linear beamformer cannot cancel all interference, especially those close in space to the speech source. To reduce the residual noise, the beamforming output can be filtered by the mask used for beamforming [136] or can be processed by a new post enhancement neural network [139] or even more iterations of neural network and beamforming [140]. However, adding a new neural network increases the size of the system, which is undesired for its deployment on hardware with limited capacities, such as hearing aids. In this chapter, we employ the exact same multi-channel DNN to predict masks for both the estimation of beamformer weights and post enhancement.

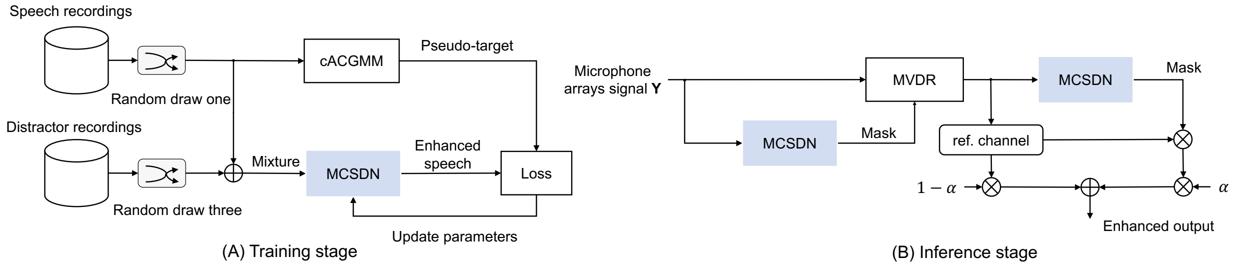


Figure 4.1: **Overview of the speech denoising system.** (A) The training stage, and (B) the inference stage. Blue blocks denote the same multi-channel speech denoising network. MVDR is a minimum variance distortionless response beamformer.

4.1.2 Method

Figure 4.1 shows an overview of the proposed method for training and inference. During training we use cACGMM to generate a better target to train a conventional multi-channel speech denoising network (MCSDN). In inference, we first apply the pretrained MCSDN, beamforming, and the same MCSDN sequentially. The final denoised signal is a weighted linear combination of the second MCSDN output and the beamforming result.

Complex Angular Central Gaussian Mixture Model

Give an M-channel recording $\mathbf{S} \in \mathbb{R}^{M\times T\times F}$ in the short-time Fourier transform (STFT) domain, where T and F denote the time frame and frequency bin, respectively, we use the complex angular central Gaussian mixture model (cACGMM) [141] to isolate the speech source $\hat{\mathbf{S}} \in \mathbb{R}^{M\times T\times F}$ and remove noisy sources from unwanted directions. cACGMM models the directional observations $\mathbf{Z}_{t,f} = \frac{\mathbf{S}_{t,f}}{||\mathbf{S}_{t,f}||}$ with a Gaussian mixture model,

$$p(\mathbf{Z}_{t,f};\Theta_f) = \sum_{k=1}^{K} \alpha_f^k \mathcal{A}(\mathbf{Z}_{t,f};\mathbf{B}_f^k), \tag{4.1}$$

where $\Theta_f = \{\alpha_f^k, \mathbf{B}_f^k \forall k\}$ denotes the model parameters, $\{\alpha_f^k \forall k\}$ is a set of k mixture weights, which are probabilities that sum to 1. $\mathcal{A}(\mathbf{s}_{t,f};\mathbf{B}_f^k)$, a complex angular central Gaussian distribution

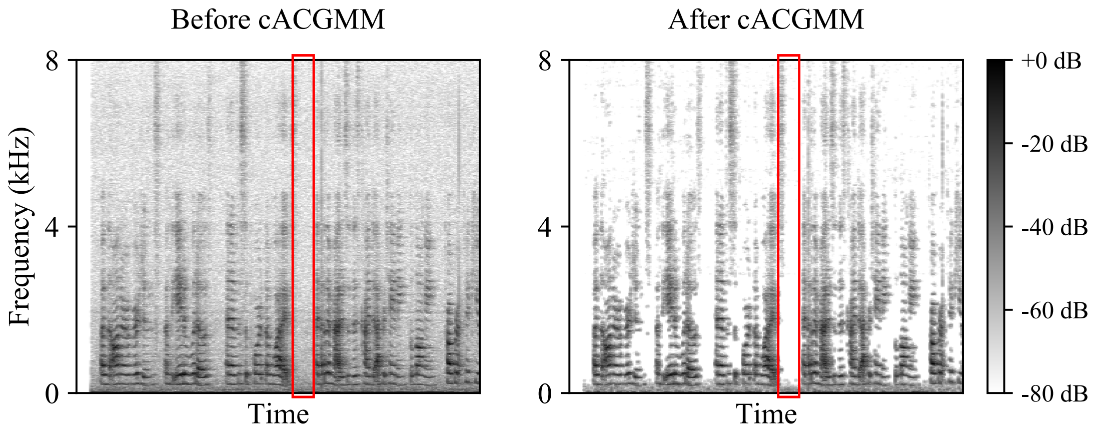


Figure 4.2: The cACGMM extracts a cleaner speech (right) signal from the recording (left) as the training target. The red rectangle highlights an instance where the background noise is attenuated.

(cACG) [142], models the distribution of $\mathbf{Z}_{t,f}$ for the component k in the mixture model as follows:

$$\mathcal{A}(\mathbf{Z}_{t,f};\mathbf{B}_f^k) = \frac{(M-1)!}{2\pi^M \det(\mathbf{B}_f^k)} \frac{1}{(\mathbf{Z}_{t,f}^H(\mathbf{B}_f^k)^{-1}\mathbf{Z}_{t,f})^M}. \tag{4.2}$$

We estimate the parameters Θ_f with the expectation-maximization (EM) algorithm. The posterior probability of $\mathbf{Z}_{t,f}$ belonging to class k is:

$$\Gamma_{t,f}^k = \frac{\alpha_f^k \mathcal{A}(\mathbf{Z}_{t,f};\mathbf{B}_f^k)}{\sum_{k=1}^K \alpha_f^k \mathcal{A}(\mathbf{Z}_{t,f};\mathbf{B}_f^k)}. \tag{4.3}$$

Since cACGMM models each frequency independently, there can be a frequency permutation problem [143], i.e., the same index k in different frequency bins point to different sources. This problem is addressed by permutation alignment [143]. Finally, we use Γ^s as the mask to extract speech,

$$\hat{\mathbf{S}} = \mathbf{S} \odot \Gamma^s, \tag{4.4}$$

where the superscript s indicates the speech, and $\odot$ denotes element-wise multiplication.

MCSDN-Beamforming-MCSDN Framework

We train a multi-channel speech denoising network (MCSDN) based on the temporal convolutional network (TCN) [144] in Conv-TasNet [25] to predict a time-frequency mask $\mathbf{M}^s \in \mathbb{R}^{T\times F}$ for the target $\hat{\mathbf{S}}$ from the multi-channel noisy signal $\mathbf{Y} \in \mathbb{C}^{M\times T\times F}$. Similar to Section 2.2, we concatenate the log power spectrogram of the reference channel signal and inter-channel phase differences (IPDs) between the reference channel and other channels as input features, since IPDs indicate which T-F bins belong to the same directional source in each frequency band. Specifically, we calculate sin(IPD) and cos(IPD) as inter-channel features. The training objective is defined as:

$$\mathcal{L} = |\mathbf{Y} \odot \mathbf{M}^s - \hat{\mathbf{S}}|. \tag{4.5}$$

We use the estimated mask $\mathbf{M}^s$ from the MCSDN for mask-based beamforming. We employ minimum variance distortionless response (MVDR) beamforming [145], which is optimized with a constraint that minimizes the power of the noise without distorting the target speech. One solution is:

$$\mathbf{w}_f = \frac{(\mathbf{\Phi}^n_f)^{-1}\mathbf{\Phi}^s_f}{\text{Trace}((\mathbf{\Phi}^n_f)^{-1}\mathbf{\Phi}^s_f)}\mathbf{u}, \tag{4.6}$$

where $\mathbf{\Phi}^n_f$ and $\mathbf{\Phi}^s_f$ are the covariance matrices of the speech and noise, respectively:

$$\mathbf{\Phi}^s_f = \frac{1}{\sum_t \mathbf{M}^s_{t,f}} \sum_t \mathbf{M}^s_{t,f}\mathbf{Y}_{t,f}\mathbf{Y}^H_{t,f}, \tag{4.7}$$

$$\mathbf{\Phi}^n_f = \frac{1}{\sum_t (1-\mathbf{M}^s_{t,f})} \sum_t (1-\mathbf{M}^s_{t,f})\mathbf{Y}_{t,f}\mathbf{Y}^H_{t,f}, \tag{4.8}$$

and $\mathbf{u}$ is a one-hot vector indicating the reference channel. H denotes conjugate transposition. Then, the linear filter $\mathbf{w}_f \in \mathbb{C}^M$ is applied to $\mathbf{Y}_{t,f} \in \mathbb{C}^M$ to generate the beamforming output $\mathbf{BF}_{t,f}$:

$$\mathbf{BF}_{t,f} = \mathbf{w}^H_f\mathbf{Y}_{t,f}. \tag{4.9}$$

Next, we use the trained MCSDN to denoise the beamforming output. To extend the single-channel beamforming output to a multi-channel one, $\mathbf{BF} \in \mathbb{C}^{M\times T\times F}$, we stack the beamforming output on each channel $[\mathbf{BF}_1, \mathbf{BF}_2, \ldots, \mathbf{BF}_m]$ by shifting the one-hot vector $\mathbf{u}$ in Equation 4.6 without introducing any new computation. Finally, the MCSDN takes $\mathbf{BF}$ as input and estimates a speech mask $\hat{\mathbf{M}}^s \in \mathbb{R}^{T\times F}$ to further denoise the beamforming output:

$$\overline{\mathbf{S}} = \hat{\mathbf{M}}^s \odot \mathbf{BF}. \tag{4.10}$$

We can view the pipeline in this way: the first time-frequency mask $\mathbf{M}^s$ is for mask-based beamforming that results in a less noisy mixture, then the second mask $\hat{\mathbf{M}}^s$ is to extract the speech from the less noisy signal. However, using the spectral mask estimated by neural networks to extract speech will inevitably cause non-linear speech distortion, which is undesirable for human listeners. We balance the noise reduction and speech distortion by mixing the beamforming output and the neural network output using a gate $\alpha \in [0, 1]$,

$$\widetilde{\mathbf{S}} = \alpha \cdot \mathbf{BF} + (1 - \alpha) \cdot \overline{\mathbf{S}}. \tag{4.11}$$

4.1.3 Experimental Settings

Data Collection

We recorded a collection of in-room speech and ambient sound samples to be used to generate sound mixtures. Each sample captures the real room acoustics, including reverb and background noise. The recording room has the dimensions 7.5 (length) x 3.5 (width) x 3 (height) meters (T60 $\approx 0.37s$). A Bruel and Kjaer Type 4128-C Head and Torso Simulator (HATS) is placed in the center of the room on a motorized turntable, which sits atop a wooden table that spans the majority of the room length-wise. For this experiment, we used two proprietary arrays of 16 microphones, one placed around each of the two ears. Surrounding the HATS are six Genelec 8020D 4" powered studio monitors for audio source playback. All speakers face towards the HATS. The speakers are placed at azimuth values ranging from 0 to 360°, and the distance from the HATS and elevation from the ground values ranges among 1, 2, or 3 meters. Once placed, the speaker locations are fixed and do not change for the given room.

Playback source data consists of the sound clips from FSD50K [146] (~50h and 11h from training and test sets) and LibriSpeech [125] (~40h and 11h from training and test sets). We resampled the playback data to 48 kHz, and pre-processsed to trim silence from the beginning and end of each clip. We manually normalized the ound clips so that the clip's db SPL at the speaker is as close as possible to a real-world example for that sound class. The target db SPL values for each sound class have a random variance of ±5 db SPL. We played each sound clip through a speaker assigned at random, and recorded through the microphone arrays at 48 kHz with 32-bit floating point precision.

Acoustic Scene Generation

For the training and development sets, we generated 12,000 and 4,000 9-second mixtures, respectively. For each mixture, we randomly draw one speech recording and three distractor recordings from the in-room recordings and mix them to simulate challenging environments. We excluded the following broad class labels from FSD50K: speech, alarm, domestic animal sounds, domestic sounds (faucet, cutlery, drawers, etc.). A clip of ambient noise recorded in the room without loudspeakers playing is also added into the mixture. The overall SNR of the mixture with respect to the speech recording varies between -3 dB and -30 dB. We resampled mixtures to 16 kHz.

Networks

We adopt the TCN module in Conv-TasNet[25] with STFT as the encoder, iSTFT as the decoder, and use 4 repeated stacks each having 6 1-D convolutional blocks in the masking network. STFTs are computed with a window size of 32 ms and a hop size of 8 ms. The effective receptive field of the model is approximately 4 s. For computational efficiency, only 8 of the channels (4 from each array) were used to train the MCSDN. Adam is employed as the optimizer with an initial learning rate of 0.001.

Evaluation

We performed two subjective evaluations: one to measure the listening improvement, and the second to evaluate the importance of cACGMM in the training. We recruited 60 self-reported normal-hearing subjects who are native English speakers to participate in a listening test on Amazon Mechanical Turk [147]. Subjects were instructed to wear headphones or earphones during the test.

The test is a simplified version of multiple stimuli with hidden reference and anchor (MUSHRA) [148]. We use 10 sets of male speaker samples and 10 sets of female speaker samples for the test. When evaluating each set of sounds, the subjects are instructed to listen to a 9-second unprocessed noisy speech sample first, and then listen to and rate the processed speech without knowing which algorithm had been applied. The processed speech came from 1) MCSDN, 2) MCSDN-Beamforming (mask-based MVDR beamforming), 3) MCSDN-Beamforming-MCSDN, 4) the remixed one with 20% from the beamforming and 80% from the second MCSDN, and 5) the target speech signal from cACGMM as shown in Equation 4.4. The subjects rate each processed speech sample on a scale with the following labels: bad (1), poor (2), fair (3), good (4), and excellent (5) on the following four aspects:

1. *Intelligibility*: How well can you recognize what the speaker is saying?
2. *Noise reduction level*: How much of the noise is removed compared to the unprocessed speech?
3. *Free of distortion*: How distortionless is the speech signal?
4. *Listening improvement*: How much the processed signal improves listening compared to the unprocessed one, e.g., how much would you like to use such a device to help them improve listening?

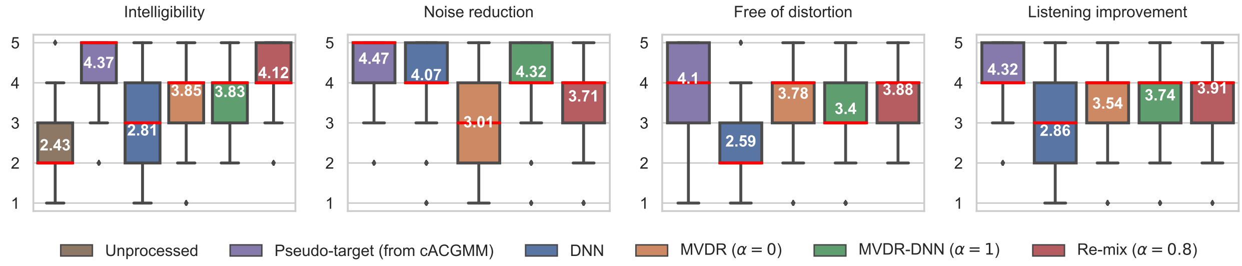


Figure 4.3: Subjective evaluation results shown as boxplots. Pseudo-target (from cACGMM) is the training target of MCSDN. The red line represents the median (all the ratings are discrete numbers from 1 to 5, so the median is discrete). The number in white denotes the mean.

Similar to MUSHRA, we used the speech signals from cACGMM as a hidden reference that were used to disqualify subjects who gave low-intelligibility and noise-reduction scores. Because the hidden reference signals were processed from individual recordings, they contained almost no noise and should have good intelligibility scores. Then, the ratings for a set of signals from a subject were disqualified and dropped if the sum of intelligibility and noise reduction scores for the hidden reference is lower than the bottom 15% of this summation from all subjects.

The setup for the second (cACGMM) experiment was similar to the first experiment. We provided two re-mix models with the MCSDNs trained with and without cACGMM, respectively, and asked the subjects to rate them for noise reduction and listening improvement.

4.1.4 Results and Discussion

Figure 4.3 compares different models in terms of the factors described above. First, all models improve the average intelligibility score over the unprocessed mixture. We notice some subjects gave high intelligibility scores to some of the unprocessed signals even when they had low SNRs. We think this is because humans can attend to a source in the presence of multiple distracting stimuli thanks to the cocktail party effect [149], thus they may focus on and exert themselves to understand the target speech. But overall, the strong background noise makes the speech much less intelligible. While the MCSDN has a high score for noise reduction due to the power of nonlinear models, it can cause speech distortion. If the mixture is too noisy, the model may also filter out speech components when it removes the noise. Therefore, the MCSDN only provides a slight

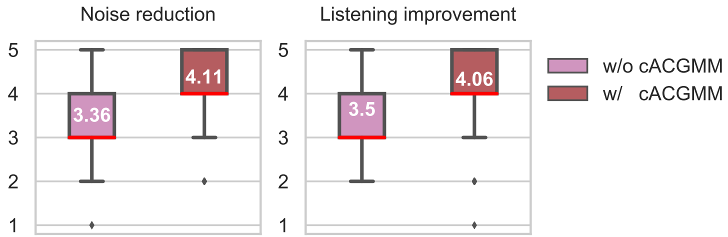


Figure 4.4: Comparison between the re-mix model using DNNs trained with cACGMM and without cACGMM. The red line represents the median (all the ratings are discrete numbers from 1 to 5, so the median is discrete). The number in white denotes the mean.

intelligibility improvement and poor listening improvement.

The MVDR beamformer uses linear filters to avoid distortion, which sacrifices the ability to cancel some noise. So, it has a lower noise-reduction score but a higher “free of distortion" score. The intelligibility and listening improvement are better than those for MCSDN.

The second MCSDN, following the beamformer, reduces the residual noise noticeably but still lifts speech distortion slightly. It does not affect intelligibility and achieves slightly better listening improvement than beamforming. All metrics at the output of the second MCSDN are significantly better than the first MCSDN.

When 20% of the beamforming output and 80% of the second MCSDN output are mixed as a new signal, we see it improves the intelligibility score over other models and achieves the best listening improvement. Some subjects mentioned they felt comfortable when the sound contained a little background noise. One explanation is that it is more realistic than over-denoised sound. Moreover, the beamforming output can mask the distorted components introduced by the neural networks.

The cACGMM target output achieves the highest mean scores in all aspects. This is expected because it is processed from moderately noisy recordings while the models’ outputs are processed from much noisier mixtures. Here, cACGMM is shown to be able to produce good quality signals to serve as the training target for MCSDN. More than 75% of the intelligibility, noise reduction, and overall listening improvement scores have a rating of at least 4. We notice the score for “free of distortion" is lower, perhaps because cACGMM estimates a probabilistic time-frequency mask through spatial clustering, which may introduce nonlinear distortion.

Figure 4.4 compares the output of the full re-mix model when the DNNs are trained with and without the cACGMM. We see the model trained using the cleaner target speech provided by cACGMM results in better noise reduction and thus better listening improvement. This demonstrates that cACGMM can help improve the quality of training data coming from real recordings.

4.2 Unsupervised Multi-Channel Separation and Adaptation

4.2.1 Introduction

In Section 4.1, we present a method that employs spatial clustering in an unsupervised manner to extract cleaner speech signals from real recordings to serve as the training target. However, the subsequent steps still adhere to supervised learning protocols. We construct mixtures by synthetically mixing several recordings. Unfortunately, this can result in a mismatch in the distribution of sound types and acoustic conditions between the simulated sound mixtures and real-world audio. For example, conversational speech is mismatched to the read speech that is typically used to train speech enhancement and separation models. Ideally, unsupervised methods alone should help to overcome the mismatch problem by directly training on real recordings from the target domain, without the need for subsequent supervised learning. However, the spatial clustering based methods [141, 150, 151, 58] including the one we use in Section 4.1, often struggle with co-located sources and strongly reverberant environments. Consequently, these unsupervised techniques are typically utilized to generate pseudo-targets, aiding further supervised training. We need more poweful unsupervised learning methods capable of handling complex mixtures effectively.

Mixture invariant training (MixIT) [61] is a recent unsupervised approach that has demonstrated competitive single-channel sound separation performance. MixIT uses mixtures of mixtures as the “noisy” input and uses the individual mixtures as weak references. The model estimates individual sound sources that can be recombined to reconstruct the original reference mixtures. Note that MixIT incurs another type of mismatch in which there are more active sources in the mixture-of-mixtures than there are in an individual mixture. Experiments have shown that when unsupervised training with MixIT and supervised training are performed jointly, the mismatch introduced by one training method is mitigated by the other. MixIT has been shown to be effective at adapting single-channel speech separation models to real-world meetings [62].

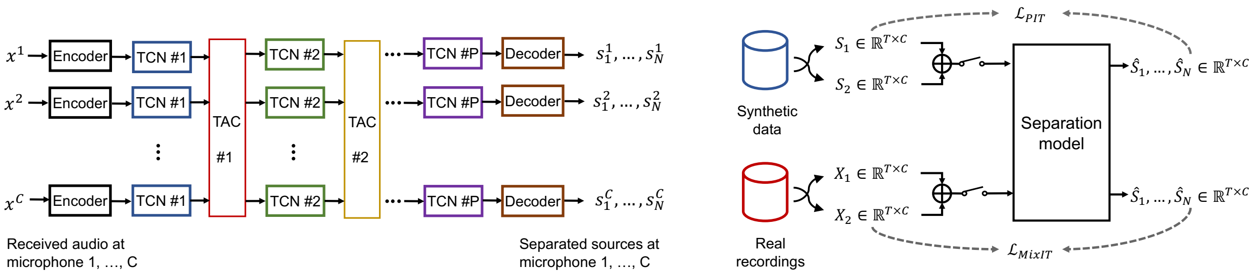
In this section, we extend MixIT to multi-channel data, allowing the model to use both spatial and spectral information to better separate sound sources. We use a separation model with multi-channel input and multi-channel output that employs a temporal convolutional network (TCN) [25, 15] and a transform-average-concatenate (TAC) module [152, 153], which enables the model to be applied to any number of microphones and any array geometries. This flexibility is particularly advantageous for models trained on diverse real-world meeting data captured by different microphone arrays. We show that when used with our flexible multi-microphone neural network, MixIT training on real mixtures improves separation and enhancement of speech on real meetings containing spontaneous speech and recorded with multiple microphones.

4.2.2 Method

Fig. 4.5(A) describes our multi-channel speech separation model. It accepts waveform input from multiple microphones and produces the multi-channel image of each source. This separation model can be trained through supervised learning with permutation invariant training (PIT) [22], through unsupervised learning with multi-channel mixture invariant training (MC-MixIT), or through a combination of both as shown in Fig. 4.5(B). We describe the extension of MixIT to MC-MixIT in this section.

Multi-Channel Speech Separation Model

The model is a variant of a single-channel separation model, TDCN++ [15]. To enable use on multi-channel audio, we interleave transform-average-concatenate (TAC) layers [152] between temporal convolutional neural networks (TCNs) [25] to exploit spatio-temporal information across channels. The model shares some similarities with VarArray [153]. Both models are designed to be invariant to microphone array geometry and the number of microphones used. However, there


x^1
x^2
x^C
Encoder
TCN #1
TAC #1
TCN #2
TAC #2
TCN #P
Decoder
$s_1^1, \dots, s_N^1$
$s_1^2, \dots, s_N^2$
$s_1^C, \dots, s_N^C$
Received audio at microphone 1, …, C
Separated sources at microphone 1, …, C
$\mathcal{L}_{PIT}$
$S_1 \in \mathbb{R}^{T \times C}$
$S_2 \in \mathbb{R}^{T \times C}$
Synthetic data
Separation model
$\hat{S}_1, \dots, \hat{S}_N \in \mathbb{R}^{T \times C}$
$X_1 \in \mathbb{R}^{T \times C}$
$X_2 \in \mathbb{R}^{T \times C}$
Real recordings
$\hat{S}_1, \dots, \hat{S}_N \in \mathbb{R}^{T \times C}$
$\mathcal{L}_{MixIT}$

(A) Multi-channel speech separation model

(B) Supervised (top) and unsupervised (bottom) learning

Figure 4.5: (A) The architecture of the proposed multi-channel input and multi-channel output speech separation model. Blocks with the same color share parameters. (B) The schematic of supervised learning with PIT on synthetic data (top) and unsupervised learning with MixIT on real recordings (bottom).

are two main differences between our model and VarArray. First, VarArray calculates a feature set from STFT coefficients, while our model takes the raw waveform directly as input. Second, VarArray merges all channels at an intermediate layer and estimates a single time-frequency mask for each source, while our model estimates a multi-mic waveform for each source.

Given a C-channel time-domain signal $\boldsymbol{X} \in \mathbb{R}^{T \times C}$, where T is the duration of the signal, we apply a linear encoder followed by a ReLU activation to transform each channel $\boldsymbol{x}^c \in \mathbb{R}^T, c = 1, \dots, C$ to a two-dimensional representation $\boldsymbol{E}^c \in \mathbb{R}^{F \times L}$, where F is the number of encoder bases and L is the number of time frames. Then, $\{\boldsymbol{E}^c\}_{c=1}^{C}$ is fed into a series of alternating TCN and TAC layers. A TCN block comprises multiple dilated convolution layers, with the output of the i^{th} TCN block in channel c denoted as $\boldsymbol{P}_i^c \in \mathbb{R}^{K \times L}$, where K is the number of features. To extract cross-channel features, we employ TAC layers that aggregate the outputs from each channel, extract cross-channel information, and feed it back to individual channels. Following the approach in [153], the output of the i-th TAC layer in channel c, denoted as $\boldsymbol{Q}_i^c \in \mathbb{R}^{2K \times L}$, is:

$$\boldsymbol{Q}_i^c = \left[\text{ReLU}(\boldsymbol{W}_i \boldsymbol{P}_i^c), \frac{1}{C} \sum_c \text{ReLU}(\boldsymbol{U}_i \boldsymbol{P}_i^c) \right], \tag{4.12}$$

where $\boldsymbol{W}_i, \boldsymbol{U}_i \in \mathbb{R}^{K \times K}$ are linear transforms. After the final TCN block, we use a sigmoid activation to predict a mask for each source in each channel and use a linear decoder to transform the masked representation back to the waveform. Note that if we remove all the TAC layers, the architecture is equivalent to independently applying a single-channel TDCN++ to individual channels. The TCN blocks process each channel locally while the TAC layers allow for inter-channel information flow.

MixIT

MixIT uses N (typically $N = 2$) reference mixtures $\boldsymbol{x}_n \in \mathbb{R}^T$, which are the columns of a matrix $\boldsymbol{X} \in \mathbb{R}^{T \times N}$. A mixture of mixtures (MoM) is formed by summing these reference mixtures to produce $\hat{\boldsymbol{x}} = \sum_N x_n$. The network then generates $M > N$ estimated sources $\left\{\boldsymbol{s}_m \in \mathbb{R}^T\right\}_{m=1}^{M}$, which are the columns of a matrix $\boldsymbol{S} \in \mathbb{R}^{T \times M}$. The MixIT loss estimates a mixing matrix $\boldsymbol{A} \in \mathbb{B}$, where $\mathbb{B} = \{0, 1\}^{M \times N}$ is a constrained set of $M \times N$ binary matrices where each row sums to 1: that is, the set of matrices which assign each estimated source $\boldsymbol{s}_m$ to one of the reference mixtures $\boldsymbol{x}_n$. Given the mixing matrix, a signal level loss, $\mathcal{L}$, measures the error between reference mixtures and their assigned estimates:

$$\mathcal{L}_{\text{MixIT}}(\boldsymbol{X}, \boldsymbol{S}) = \min_{A \in \mathbb{B}} \mathcal{L}(\boldsymbol{X}, \boldsymbol{S}\boldsymbol{A}). \tag{4.13}$$

where $\mathcal{L}$ typically operates column-wise, so that $\mathcal{L}(\boldsymbol{X}, \boldsymbol{S}) = \sum_n \mathcal{L}(\boldsymbol{x}_n, (\boldsymbol{S}\boldsymbol{A})_n)$. In this study, $\mathcal{L}$ is negative thresholded SNR:

$$\mathcal{L}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 10 \log_{10} \left(\frac{||\boldsymbol{y}||^2}{||\hat{\boldsymbol{y}} - \boldsymbol{y}||^2 + \tau ||\boldsymbol{y}||^2} \right), \tag{4.14}$$

where τ is a soft limit on the maximum SNR. We select $\tau = 0.001$.

Multi-Channel MixIT

We extend MixIT by applying the same mixing matrix to all the channels of each source. We write the multi-channel references $\boldsymbol{X}^c$ and sources $\boldsymbol{S}^c$ at channel c. The MixIT loss requires finding the optimal mixing matrix $\boldsymbol{A} \in \mathbb{B}$ across all channels:

$$\mathcal{L}_{\text{MC-MixIT}}(\{\boldsymbol{X}^c\}, \{\boldsymbol{S}^c\}) = \min_{A \in \mathbb{B}} \sum_c \mathcal{L}(\boldsymbol{X}^c, \boldsymbol{S}^c \boldsymbol{A}). \tag{4.15}$$

By sharing $\boldsymbol{A}$ across microphones, the loss encourages the order of the separated sources in the model outputs to be consistent across all channels.

4.2.3 Experimental Settings

Our experimental approach largely follows that of Sivaraman et al. [62]. We conducted experiments using the AMI Corpus [121] of meeting room recordings for evaluation data and as one source of training data. Hyperparameters were fixed to values in Table 4.1 across all experiments; only training procedures were varied.

Table 4.1: Hyperparameter values.

Category	Hyperparameter	value
Model	TCN superblocks	4
	TCN blocks per superblock	8
	TCN kernel width	3
	TCN window size	64 samples
	TCN hop size	32 samples
	TCN bottleneck dim	128
	TCN conv channels	512
	TAC projection dim	128
	# of output sources	8
Data	Unsup example len	10 seconds
	Sup example len	5 seconds
	Audio sample rate	16 kHz
Training	Trainable weights	4.7 million
	Optimizer	Adam
	Batch size	256
	Learning rate	$3 * 10^{-4}$
	Training steps	1 million

Training

All models have $M = 8$ output sources, though depending on the training configuration, there may be fewer target sources. When there are fewer than 8 target sources, our negative thresholded SNR loss is applied to only the non-zero sources, and we rely on mixture consistency [154] to push unused outputs to zero.

Our training configurations are shown in Figure 4.5 (B). For unsupervised (MixIT) training, we use mixtures of randomly chosen segments from the AMI Corpus. Our AMI training split is 71 hours.

For supervised (PIT) training, we use single-talker segments of the AMI Corpus to synthesize training mixtures where each source has only a single active talker, as identified using the AMI annotations. To generate each training example, two such segments are taken from the same room but from different speakers, who we refer to as speaker 1 and speaker 2. These are used as references that are added together to create a synthetic input mixture. We experiment with two different approaches for constructing references.

The first approach addresses this problem by creating a cleaner reference following the procedure described in Sivaraman et al. [62] to create "synthetic overlapping AMI" (referred to as "synth AMI" in Table 4.2 and Table 4.3). To define a nearly noise-free speech reference, we use the headset mic recording for speaker 2's segment and find the multi-channel filter that optimally matches the microphone array signal. We define that multi-channel filtered headset signal to be the training target for speaker 2. We define the training target for speaker 1 to be the microphone array recording of speaker 1's segment (including background noise), and we define a third training target that is the residual from speaker 2's microphone array recording after subtracting the filtered headset target. These three targets add up to the sum of the two microphone array segments, but they are asymmetric with respect to the speakers. Because the speaker 2 reference is cleaner, we focus on speaker 2 during evaluation, below. See Sivaraman et al. [62] for additional details.

In the second approach, we directly use the array signals for speaker 1 and speaker 2 as references; we refer to this as "mixed AMI" in Table 4.3. A caveat with this approach is that both reference signals contain some background noise, and thus the mixture contains double background noise. As training targets, the noisy references may lead the model to preserve noise in its speech estimates. For evaluation these references may not be as accurate as desired.

For 1-microphone training, we use only the first channel of the AMI recordings. For 2-microphone, 4-microphone, and 8-microphone training, we use 2, 4, and 8 microphones from the circular table-mounted microphone array at 180°, 90°, and 45° separation from one another, respectively. For all model evaluation, we evaluate only the first channel of the output. Train, validation, and test splits follow the standard AMI "full-corpus" partition of meetings.

As a baseline and as a model from which to warm-start, we train a single-channel model on 1600 hours of audio from videos in the YFCC100M corpus with train, validation, and test splits from Wisdom et al. [61].

All models are trained with a mixture consistency hard constraint [154] on their outputs and with feature-wise layer normalization as described in Kavalerov et al. [15]. We train all models for one million steps.

Evaluation

To objectively evaluate our methods on synthetic AMI, we use scale-invariant signal-to-noise ratio improvement (SI-SNRi) with the filtered headset signal as reference [155]. To measure subjective audio quality, we use the multiple stimulus with hidden reference and anchors (MUSHRA) [148] for the synthetic AMI evaluation dataset, with the filtered headset as reference. For real AMI data, where we do not have a known clean reference, we adopt a variant of MUSHRA that allows for imperfect references called MUSHIRA [62]. For MUSHIRA on real AMI data, the imperfect reference is a headset recording of a target speaker that contains cross-talk. For all listening tests, audio is presented diotically, with the signal corresponding to the first microphone being presented to both ears. We collect 5 ratings per example for both MUSHRA and MUSHIRA.

We trained two instances of each configuration and report averages across model instances for all metrics in both Table 4.2 and Table 4.3.

Table 4.2: Cross-evaluation by number of mics. Values are SI-SNRi in dB. “S1” and “S2” refer to the full-duration speaker and the overlapping speaker, respectively. Due to space constraints, we report results on only “synth AMI” training data without warm start.

# of mics	Training	1-mic eval		2-mic eval		4-mic eval		8-mic eval	
- training	method	S1	S2	S1	S2	S1	S2	S1	S2
1-mic	Sup	4.3	6.0	4.5	6.5	4.5	6.5	4.5	6.6
	Unsup	3.7	10.0	3.6	9.9	3.6	9.9	3.6	9.9
	Semi	**6.0**	**10.3**	6.1	10.6	6.1	10.6	6.1	10.6
2-mic	Sup	1.8	-3.5	5.8	7.9	6.1	8.4	6.2	8.5
	Unsup	3.2	8.6	4.6	**11.8**	4.7	11.7	4.8	11.9
	Semi	4.3	5.5	**6.8**	**11.8**	6.8	12.0	6.9	12.1
4-mic	Sup	0.5	-7.3	4.8	6.1	6.3	9.7	6.3	9.7
	Unsup	1.5	3.1	4.2	10.9	4.8	12.8	4.9	13.0
	Semi	2.6	2.3	5.9	10.5	**7.0**	**12.8**	7.0	12.8
8-mic	Sup	-0.8	-8.6	3.8	4.4	5.8	9.9	6.3	10.6
	Unsup	1.3	3.0	4.3	10.9	4.9	12.9	5.0	**13.6**
	Semi	2.5	3.4	5.6	9.5	6.6	11.4	**7.1**	11.2

4.2.4 Results and Discussion

Our TCN-TAC architecture allows models trained with any number of input microphones to be applied to data with a different number of microphones. Table 4.2 cross-evaluates models trained on N microphones on the evaluation sets for M microphones, for $N, M \in [1, 2, 4, 8]$. We observe that 1-mic trained models perform nearly the same no matter how many input mics are provided. For $N > 1$, quality is best when the number of training microphones equals the number of input microphones, but models degrade gracefully when given a different number of microphones, and in some cases quality improves modestly when additional mic inputs are provided beyond what was used for training. We also observe that unsupervised learning outperforms supervised learning in most cases on speaker 2 (which we focus on because its reference is cleaner as described in Section 4.2.3), and combining both training methods can further improve the performance.

Due to human evaluation capacity constraints, we were unable to do human listening eval of all models. Instead, in Table 4.3, we take 1-microphone and 4-microphone models as examples and provide a comprehensive comparison of different model training configurations in terms of

Table 4.3: AMI data results. “S1” and “S2” refer to SI-SNRi for the full-duration speaker and the overlapping speaker, respectively. The first microphone is used as the reference for reference-based metrics. For full synthetic AMI, the absolute input SI-SNRs are 0.5 dB for S1 and -9.2 dB for S2, which are used in the SI-SNRi computation. “Warm” indicates loading the model weight pre-trained with MixIT on 1600 hours of YFCC100M data (single-channel). The pooled 95% confidence intervals are ±1.1 for the MUSHRA and ±2.2 for the MUSHIRA ratings.

Model Configuration			Synthetic AMI			Real AMI
Sup PIT	Unsup MixIT	Warm	S1	S2	MUSHRA	MUSHIRA
Baselines						
Headset			–	–	96.6	93.4
Headset filtered to distant mic			∞	∞	64.5	50.0
Distant mic			0.0	0.0	33.1	38.8
1-microphone						
–	YFCC	–	2.1	2.5	29.1	29.8
–	AMI	–	3.6	10.0	38.4	**43.5**
Mixed AMI	–	–	-1.0	6.7	39.4	38.9
Synth AMI	–	–	4.3	6.0	35.8	40.9
Mixed AMI	AMI	–	0.0	10.2	41.8	39.8
Synth AMI	AMI	–	6.0	10.3	39.0	41.8
–	AMI	YFCC	3.7	9.8	40.4	41.1
Mixed AMI	AMI	YFCC	-0.4	9.4	42.1	42.6
Synth AMI	AMI	YFCC	**6.4**	**14.1**	**42.9**	41.7
4-microphone						
–	AMI	–	4.8	12.8	43.7	44.3
Mixed AMI	–	–	0.4	10.9	43.9	43.8
Synth AMI	–	–	6.3	9.7	37.5	38.2
Mixed AMI	AMI	–	1.9	12.5	44.7	43.9
Synth AMI	AMI	–	7.0	12.8	40.9	**46.2**
–	AMI	YFCC	4.5	12.0	44.2	44.3
Mixed AMI	AMI	YFCC	0.2	12.5	43.8	42.7
Synth AMI	AMI	YFCC	**7.2**	**16.4**	**46.5**	46.1

SI-SNRi scores on the fully synthetic AMI evaluation dataset, MUSHRA scores for a subset of a few hundred synthetic AMI examples, and MUSHIRA scores for about 100 real overlapping AMI examples.

In the 1-microphone subtable, unsupervised training with MixIT on AMI outperforms supervised training on either mixed AMI or synthetic AMI across most metrics. However, it falls short in terms of the MUSHRA score compared to supervised training on mixed AMI (38.4 v.s. 39.4), and in terms of SI-SNRi for speaker 1 compared to supervised training on synthetic AMI (3.6 v.s.

4.3). Combined supervised and unsupervised training further improves the SI-SNRi and MUSHRA scores on synthetic AMI, but does not improve the MUSHIRA score on real AMI.

The model trained with MixIT on YFCC100M performs quite poorly on the AMI eval set across all metrics. However, using the YFCC100M-trained model to warm-start can further improve MUSHRA and MUSHIRA scores when using mixed AMI as the supervised dataset, while improving SI-SNRi and MUSHRA scores when using synthetic AMI as the supervised dataset.

Each 1-microphone model has a 4-microphone counterpart, with the exception of the single-channel separation model trained on YFCC100M, which is evaluated alongside other models and also used to warm-start some other configurations. Each pair of 1-microphone and 4-microphone models share the same learning strategies, with the only difference being that the 4-microphone models take advantage of multi-channel signals by using TAC modules to exploit spatial features for better separation. (TAC modules are present in both 1- and 4-microphone models, but they only have the effect of transferring information across channels in the 4-microphone models.) Notice that all 4-microphone models significantly outperform their 1-microphone counterparts in terms of SI-SNRi, MUSHRA, and MUSHIRA scores, except for the 4-microphone model trained supervised on synthetic AMI, which had a lower MUSHIRA score (38.2) compared to 1-microphone (42.9).

Overall, the results demonstrate that multi-channel models achieve better separation performance in supervised learning, unsupervised learning, and their combination. This confirms that multi-channel models can also take advantage of unsupervised learning to adapt on real-world multi-channel recordings. Notably, when training the multi-channel model using both MC-MixIT unsupervised and PIT supervised on synthetic AMI, warm-starting from a model pre-trained with MixIT on monaural YFCC100M achieves significant improvement across all synthetic AMI eval metrics. It is likely YFMCC100m contains rich acoustics including diverse speakers and environmental conditions. In the pre-training stage, the model focuses on separating sources using solely spectral-temporal information and ends with an effective model weight initialization which benefits the model further exploring spatial information to improve separation in the next stage.

Warm-starting a 4-microphone model using a 1-microphone model is feasible thanks to the architecture's invariance to the number of microphones. This highlights the potential of pre-training models on a large amount of general audio data that contains a wide variety of real-world speech and then adapting these models on a smaller number of domain-specific speech recordings from multi-mic arrays.

Finally, the most effective configuration for achieving optimal performance is the multi-microphone model that is pre-trained on YFCC100M and then employs semi-supervised training with PIT on synthetic AMI and MixIT on real AMI.

4.3 Conclusion

This chapter investigates using unsupervised learning methods to mitigate the data mismatch problem in speech separation. We explored two different directions. In the first direction, we recorded real-world audio to capture the complex acoustics of real environments that synthetic audio struggles to replicate. We mixed individual recordings with reverberation and moderate noise into a mixture with multiple distractors. Instead of using speech recordings as the learning target, we applied cACGMM on individual recordings to extract clean speech signals to serve as the target for learning, which significantly improves the training dataset. Experiments show that the supervised model trained using the cleaner target speech provided by cACGMM results in better noise reduction and thus better listening improvement.

In the second direction, we generalized a single-channel unsupervised learning method, MixIT, to multi-channel settings. We show that multi-channel MixIT enables model adaptation on real-world multi-channel unlabeled spontaneous speech recordings. Our best-performing system combines pre-training with MixIT on a large amount of single-channel data from YFCC100M, supervised training with PIT on synthetic multi-channel data, and unsupervised training with MixIT on multi-channel target domain data. In the future, we plan to explore integrating two directions and to investigate using larger and more diverse amounts of open-domain data to improve separation performance.

Part II

Brain-Controlled Hearing

Chapter 5: Speaker-Independent Auditory Attention Decoding Without Access to Clean Speech Sources

Speech perception in crowded environments is challenging for hearing-impaired listeners. Assistive hearing devices cannot lower interfering speakers without knowing which speaker the listener is focusing on. One possible solution is auditory attention decoding (AAD) in which the brainwaves of listeners are compared with sound sources to determine the attended source, which can then be amplified to facilitate hearing. In realistic situations, however, only mixed audio is available. This chapter addresses this major obstacle in actualization of AAD by using the speech separation model introduced in Chapter 2.1 to automatically separate speakers in mixed audio. Our results show that AAD with automatically separated speakers is as accurate and fast as using clean speech sounds. The integration of speech separation with AAD significantly improves the subjective and objective quality of the attended speaker.

5.1 Introduction

Speech communication in acoustic environments with more than one speaker is extremely challenging for hearing-impaired listeners [156]. Assistive hearing devices have seen substantial progress in suppressing background noises that are acoustically different from speech [9, 10], but they cannot enhance a target speaker without knowing which speaker the listener is conversing with [11]. Recent discoveries of the properties of speech representation in the human auditory cortex have shown an enhanced representation of the attended speaker relative to unattended sources [12]. These findings have motivated the prospect of a brain-controlled assistive hearing device to constantly monitor the brainwaves of a listener and compare them with sound sources in the environment to determine the most likely talker that a subject is attending to [14]. This process is termed auditory attention decoding (AAD), a research area that has seen considerable growth in recent years. Then, this device can amplify the attended speaker relative to others to facilitate hearing that speaker in a crowd.

Multiple challenging problems, including nonintrusive methods for neural data acquisition and optimal decoding methods for accurate and rapid detection of attentional focus, must be resolved to realize a brain-controlled assistive hearing device. In addition, we have only a mixture of sound sources in realistic situations that can be recorded with one or more microphones. Because the attentional focus of the subject is determined by comparing the brainwaves of the listener with each sound source, a practical AAD system needs to automatically separate the sound sources in the environment to detect the attended source and subsequently amplify it. One solution that has been proposed to address this problem is beamforming [157]; in this process, neural signals are used to steer a beamformer to amplify the sounds arriving from the location of the target speaker [158, 159]. However, this approach requires multiple microphones and can be beneficial only when ample spatial separation exists between the target and interfering speakers. An alternative and possibly complementary method is to leverage the recent success in automatic speech separation algorithms that use deep neural network models [79, 80]. In one such approach, neural networks were trained to separate a pretrained, closed set of speakers from mixed audio [160]. Next, separated speakers were compared with neural responses to determine the attended speaker, who was then amplified and added to the mixture. Although this method can help a subject interact with known speakers, such as family members, this approach is limited in generalization to new, unseen speakers, making it ineffective if the subject converses with a new person, in addition to the difficulty of scaling up to a large number of speakers. To alleviate this limitation, in Section 2.1, we have proposed a causal, speaker-independent single-channel speech separation model, online deep attractor network (ODAN), that can generalize to unseen speakers, meaning that the separation of speakers can be performed without any prior training on target speakers.

In this chapter, we address the problem of speaker-independent AAD without access to clean sources by using ODAN to automatically separate unseen sources. Because this system can generalize to new speakers, it overcomes a major limitation of the previous AAD approach that required training on the target speakers [20]. The AAD framework enhances the subjective and objective quality of perceiving the attended speaker in a multi-talker (M-T) mixture. By combining recent advances in automatic speech processing and brain-computer interfaces, this chapter represents a major advancement toward solving one of the most difficult barriers in actualizing AAD. This solution can help people with hearing impairment communicate more easily.

5.2 Materials and Methods

5.2.1 Participants and Neural Recordings

Three subjects who were undergoing clinical treatment for epilepsy at North Shore University Hospital contributed to the data described in this chapter. All patients provided informed consent as monitored by the local institutional review board and in accordance with the ethical standards of the Declaration of Helsinki. The decision to implant the electrode targets and the duration of implantation were made entirely on clinical grounds without reference to this investigation. Patients were informed that participation in this study would not alter their clinical treatment and that they could withdraw at any time without jeopardizing their clinical care. All subjects had self-reported normal hearing.

Two subjects (subjects 1 and 2) were implanted with high-density subdural electrocorticography (ECoG) arrays over their language dominant temporal lobe, providing coverage of the superior temporal gyrus (STG), which selectively represents attended speech [12]. The third subject (subject 3) was implanted with bilateral stereoelectroencephalography (sEEG), with depth electrodes in Heschl's gyrus (containing primary auditory cortex) and STG. This implantation resulted in varying amounts of coverage over the left and right auditory cortices of each subject. Figure 5.1 plotted the electrodes on the average Freesurfer brain template [161].

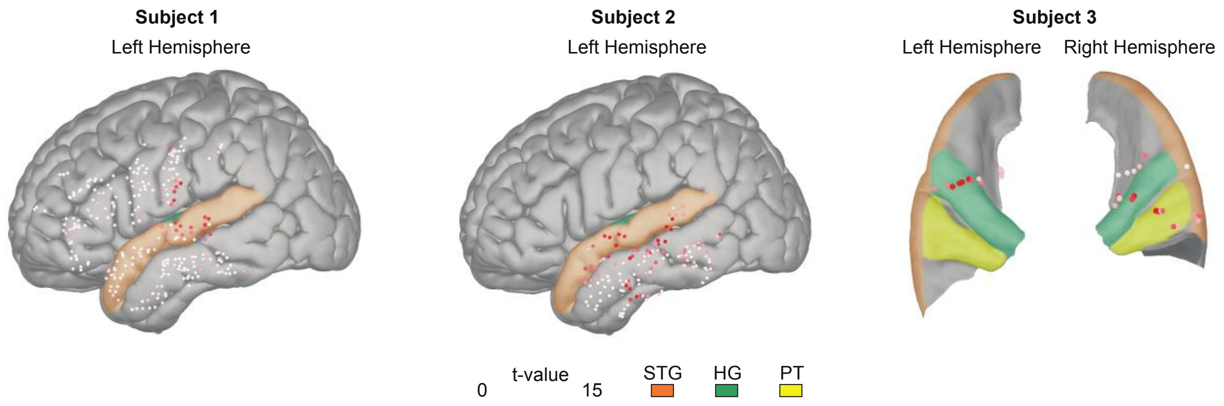


Figure 5.1: **Electrode coverage and speech responsiveness for each subject.** Subjects 1 and 2 were implanted with high-density subdural electrode arrays over their left (language dominant) temporal lobe with coverage over the superior temporal gyrus (STG; orange). Subject 3 partook in stereotactic EEG (sEEG) in which they were implanted bilaterally with depth electrodes. These differences in implantation resulted in varying coverage of the STG, Heschl's gyrus (HG; green) and planum temporale (PT; yellow) in the left and right auditory cortices. The t-value resulting from *t*-test between speech versus silence is plotted on a red color scale.

5.2.2 Data Preprocessing and Hardware

Neural data were recorded using Tucker Davis Technologies hardware and sampled at 2441 Hz. The data were resampled to 500 Hz. A first-order Butterworth high-pass filter with a cutoff frequency at 1 Hz was used to remove DC drift. Data were subsequently re-referenced using a local scheme, whereby the average voltage from the nearest neighbors was subtracted from each electrode. Line noise at 60 Hz and its harmonics (up to 240 Hz) were removed using second-order infinite impulse response (IIR) notch filters with a bandwidth of 1 Hz. A period of silence was recorded before each experiment, and the corresponding data were normalized by subtracting the mean and dividing by the SD of this prestimulus period.

Next, neural data were filtered into the high-gamma band (70 to 150 Hz); the power of this band is modulated by auditory stimuli [12, 162, 163]. To obtain the power of this broad band, we first filtered the data into eight frequency bands between 70 and 150 Hz with increasing bandwidth using Chebyshev type 2 filters. Then, the power (analytic amplitude) of each band was obtained using a Hilbert transform. We took the average of all eight frequency bands as the total power of the high-gamma band.

5.2.3 Stimuli and Experimental Design

Each subject participated in the following experiments for this study: single-talker (S-T) and multi-talker (M-T) experiments. In the S-T experiment, each subject listened to four continuous speech stories (each story was 3 min long), for a total of 12 min of speech material. The stories were uttered once by a female and once by a male speaker (hereafter referred to as Spk1 and Spk2, respectively). For the M-T experiment, the subjects were presented with a mixture of the same speech stories as those in the S-T experiment, where both speakers were combined at a 0-dB target-to-masker ratio. The M-T experiment was divided into four behavioral blocks, each containing a mixture of two different stories spoken by Spk1 and Spk2. Before each experimental block, the subjects were instructed to focus their attention on one speaker and to ignore the other. All the subjects began the experiment by attending to the male speaker and switched their attention to the alternate speaker on each subsequent block. To ensure that the subjects were engaged in the task, we intermittently paused the stories and asked the subjects to repeat the last sentence of the attended speaker before the pause. All the subjects performed the task with high behavioral accuracy and were able to report the sentence before the pause with an average accuracy of 90.5% (S1, 94%; S2, 87.%; and S3, 90%). Speech sounds were presented using a single Bose SoundLink Mini 2 loudspeaker placed in front of the subject at a comfortable hearing level, with no spatial separation between the competing speakers.

5.2.4 Speaker-Independent AAD

Figure 1.1 shows a schematic of the proposed speaker-independent AAD framework. A speaker separation algorithm first separates the speakers in M-T mixed audio. Next, the spectrograms of the separated speakers are compared with the spectrogram that is reconstructed from the evoked neural responses in the auditory cortex of the listener to determine the attended speaker. Then, the attended speaker is amplified relative to other speakers in the mixture before it is delivered to the listener.

In this chapter, we employed the online Deep Attractor Network (ODAN), as introduced in Section 2.1, for speaker separation. ODAN, a causal and speaker-independent automatic speech separation algorithm, is capable of generalizing to speakers it has not seen before. This feature allows the AAD system to access individual speakers' speech without prior information about them. We refer to this integrated setup as the ODAN-AAD system. To compare separated speech with clean individual speech in AAD, we also considered an ideal scenario where we have access to individual clean speech, which is impractical without speech separation in real-life situations. We label this system as the Clean-AAD system.

5.2.5 Stimulus Reconstruction

To determine the attended speaker, we used a method known as stimulus reconstruction [164, 165]. This method applies a spatiotemporal filter (decoder) to neural recordings to reconstruct an estimate of the spectrogram that a user is listening to. The decoder is trained by performing linear regression to find a mapping between the neural recordings and spectrogram. We trained the decoders on S-T data to minimize any potential bias that may result from training the decoders on the M-T data. After we trained the decoders using S-T data, we used the same decoders to reconstruct spectrograms from the M-T experiment [12].

The decoders were trained using the electrodes that were significantly more responsive to speech than to silence. To perform these statistical analyses, we segmented the neural data into 500-ms chunks and divided them into the following categories: speech and silence. Significance was determined using unpaired *t*-test (false discovery rate corrected, $q < 0.05$). This electrode selection resulted in varying numbers of electrodes for each subject (see Fig. 5.1). The decoders were trained with time lags from -400 to 0 ms. See [164] for further information on the stimulus reconstruction algorithm.

5.2.6 Decoding Accuracy

Determining to whom the subject is attending requires correlation analysis, commonly using Pearson's r value [14, 70]. Typically, the spectrogram that has the largest correlation with the reconstructed spectrogram is considered the attended speaker. We used window sizes ranging from 2 to 32 s to calculate correlations (in logarithmically increasing sizes). We defined decoding accuracy as the percentage of the segments in which the reconstructions had a larger correlation with the attended spectrogram than with the unattended spectrogram.

5.2.7 Dynamic Switching of Attention

To simulate a dynamic scenario in which a subject was switching attention between two speakers, we divided and concatenated the neural data into consecutive segments in which subjects were attending to either speaker. Specifically, we divided the data into 10 segments, each lasting 60 s. Subjects attended to the male speaker for the first segment. To assess our ability to track the attentional focus of each subject, we used a sliding window approach whereby we obtained correlation values every second over a specified window. We used window sizes ranging from 2 to 32 s (in logarithmically increasing sizes). Larger windows should lead to more consistent (less noisy) correlation values, thus providing a better estimate of the attended speaker. However, this approach should also be slower at detecting a switch in attention, therefore leading to a reduction in decoding speed.

5.2.8 Psychoacoustic Experiment

To test if the difficulty of attending to the target speaker is reduced using the ODAN-AAD system, we performed a psychoacoustic experiment on 20 healthy controls using Amazon Mechanical Turk (www.MTurk.com). The stimuli used for this experiment were the same as those used for the neural experiment, i.e., subjects were always presented with a mixture of Spk1 and Spk2. However, for ODAN-AAD and clean-AAD systems, the decoded target speaker was amplified by 12 dB. This particular amplification level has been shown to significantly increase the intelligibility of the attended speaker while keeping the unattended speakers audible enough to enable attention switching [166].

The experiment was divided into six blocks, each containing nine trials. Each trial consisted of a single sentence. One-third of the trials consisted of the raw mixture, another third contained modified audio using the ODAN-AAD framework, and the remaining third contained modified audio using the clean-AAD system. The trial order was randomized. Before each block, the subjects were instructed to pay attention to one of the speakers. After each trial (sentence), we asked the subjects to indicate the difficulty they had in understanding the attended speaker on a scale of 1 to 5 as follows: very difficult [156], difficult, not difficult, easy, and very easy [12]. From these responses, we calculated the mean opinion score (MOS) [167]. In total, the experiment lasted approximately 15 min.

5.3 Results

5.3.1 Reconstruction of the Attended Speaker from Evoked Neural Activity

The reconstructed spectrogram from the auditory cortical responses of a listener in an M-T speech perception task is more similar to the spectrogram of the attended speaker than that of the unattended speaker [12]. Therefore, the comparison of the neurally reconstructed spectrogram with the spectrograms of individual speakers in a mixture can determine the attentional focus of the listener [14]. We used a linear reconstruction method to convert neural responses back to the spectrogram of the sound (see Section 5.2.5).

To examine the similarity of the reconstructed spectrograms from the neural responses to the spectrograms of the attended and unattended speakers, we measured the correlation coefficient (Pearson's r) between the reconstructed spectrograms with both ODAN and the actual clean spectrograms of the two speakers. The correlation values were estimated over the entire duration of the M-T experiment. As shown in Fig. 5.2C, the correlation between the reconstructed and clean spectrograms was significantly higher for the attended speaker than for the unattended speaker (paired t-test, $P < 0.001$; Cohen's $D = 0.8$). This observation shows the expected attentional modulation

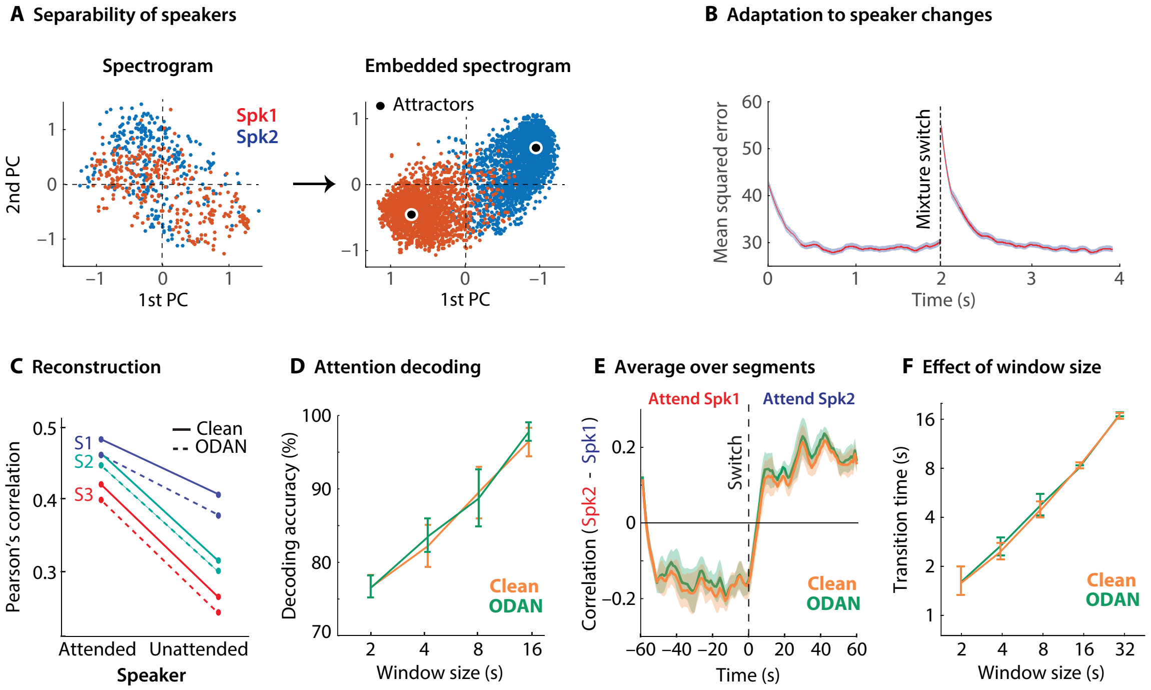


Figure 5.2: **Evaluating the accuracy of speech separation and attention decoding methods.** (A) Comparison of separation between the representation of the two speakers in the T-F (left) and embedding space (right). The axis represents the first two principal components of the data that are used to allow visualization. Each dot represents one T-F bin (left) or one embedded T-F bin (right), which are colored based on the relative power of the two speakers in that bin. (B) Separation accuracy as a function of time. The dashed line shows the time at which the speakers in the mixture are switched. (C) Correlation values between the reconstructed spectrograms (from neural data) and the attended/unattended spectrograms. Correlation values were significantly higher for the attended speaker (paired *t*-test, $P < 0.001$; Cohen's $D = 0.8$), thus confirming the effect of attention in the neural data. The correlation with the clean spectrograms was slightly higher than that with the ODAN outputs, but the differences between the attended and unattended speakers were the same for both clean and ODAN outputs. (D) Attention decoding: The percentage of segments in which the attended speaker was correctly identified for a varying number of correlation window lengths when using ODAN and the actual clean spectrograms. There was no significant difference between using the clean and the ODAN spectrograms (Wilcoxon rank sum test, $P = 0.9$). (E) Dynamic switching of attention was simulated by segmenting and concatenating the neural data into alternating 60-s bins. The dashed line indicates switching attention. The average correlation values from one subject are shown using a 4-s window size for both ODAN and the actual clean spectrograms. The shaded regions denote SE. (F) The transition time in detecting a switch of attention was calculated as the time at which the correlation difference between the two speakers crossed zero. The average transition time across subjects increased with larger window sizes; however, there was no significant difference between the transition time of ODAN and the actual clean spectrograms (Wilcoxon rank sum test, $P > 0.6$).

of the auditory cortical responses [12]. The comparison of the correlation values of ODAN and the actual clean spectrograms (Fig. 5.2C) shows a similar difference value between the attended and unattended spectrograms (average correlation difference for clean = 0.125 and for ODAN = 0.128), suggesting that ODAN spectrograms can be equally effective for attention decoding. Figure 5.2C also shows a small but significant decrease in the correlation values of the reconstructed spectrograms with ODAN compared with those of the actual clean spectrograms. This decrease is caused by the imperfect speech separation performed by the ODAN algorithm. Nevertheless, this difference is small and equally present for both attended and unattended speakers. Therefore, this difference did not significantly affect the decoding accuracy as shown below.

5.3.2 Decoding the Attentional Focus of the Listener

To study how the observed reconstruction accuracy with attended and unattended speakers (Fig. 5.2C) translates into attention decoding accuracy, we used a simple classification scheme in which we computed the correlation between the reconstructed spectrograms with both clean attended and unattended speaker spectrograms over a specified duration. Next, the attended speaker is determined as the speaker with a higher correlation value. The duration of the signal used for the calculation of the correlation is an important parameter and affects both the decoding accuracy and speed (see Section 5.2.6). Longer durations increase the reliability of the correlation values, hence improving the decoding accuracy. This phenomenon is shown in Fig. 5.2D, where the varying duration of the temporal window was used to determine the attended speaker. The accuracy in Fig. 5.2D indicates the percentage of segments for which the attended speaker was correctly decoded. The accuracy was calculated for the following cases: when using ODAN spectrograms and when using the actual clean spectrograms. We found no significant difference in decoding accuracy with ODAN or the clean spectrograms when different time windows were used (Wilcoxon rank sum test, P = 0.9). This finding confirms that automatically separated sources by the ODAN algorithm result in the same attention decoding accuracy as that with the actual clean spectrograms. As expected, increasing the correlation window resulted in improved decoding accuracy for both

ODAN and actual clean sources (Fig. 5.2D).

Next, we examined the temporal properties of attention decoding when ODAN and the actual clean spectrograms were used. We simulated a dynamic switching of attention where the neural responses were concatenated from different attention experiment blocks such that the neural data alternated between attending to the two speakers. To accomplish this, we first divided the neural data in each experiment block into 60-s segments (total of 12 segments) and interleaved segments from the two attention conditions (see Section 5.2.7). We compared the correlation values between the reconstructed spectrograms with both ODAN and the actual clean spectrograms using a sliding window of 4 s. Then, we averaged the correlation values over the segments by aligning them according to the time of the attention switch. Figure 5.2E shows the average correlation for one example subject over all the segments where the subject was attending to Spk1 in the first 60 s and switched to Spk2 afterward. The overlap between the correlation plots calculated from ODAN and the actual clean spectrograms shows that the temporal properties of attention decoding are the same in both cases; hence, ODAN outputs can replace the clean spectrograms without any significant decrease in decoding speed. We quantified the decoding speed using the transition time, which is the time it takes to detect a switch in the listener's attention. Transition times were calculated as the time at which the average correlation crossed the zero line. Figure 5.2F shows the average transition times for the three subjects for five different sliding window durations. As expected, the transition times increase for longer window lengths, but there were no significant differences between ODAN and the clean spectrograms (paired *t*-test, $P > 0.7$; Fig. 5.2F).

5.3.3 Increased Subjective and Objective Perceived Quality of the Attended Speaker

We performed a psychoacoustic experiment (see Section 5.2.8) comparing the original mixture and the enhanced sounds from the ODAN-AAD system and the enhanced sounds from the clean-AAD system. The bar plots in Fig. 5.3A show the median MOS ± standard error (SE) for each of the three conditions. The average subjective score for the ODAN-AAD shows a significant improvement over the mixture (56% improvement; paired *t*-test, $P < 0.001$), demonstrating that the

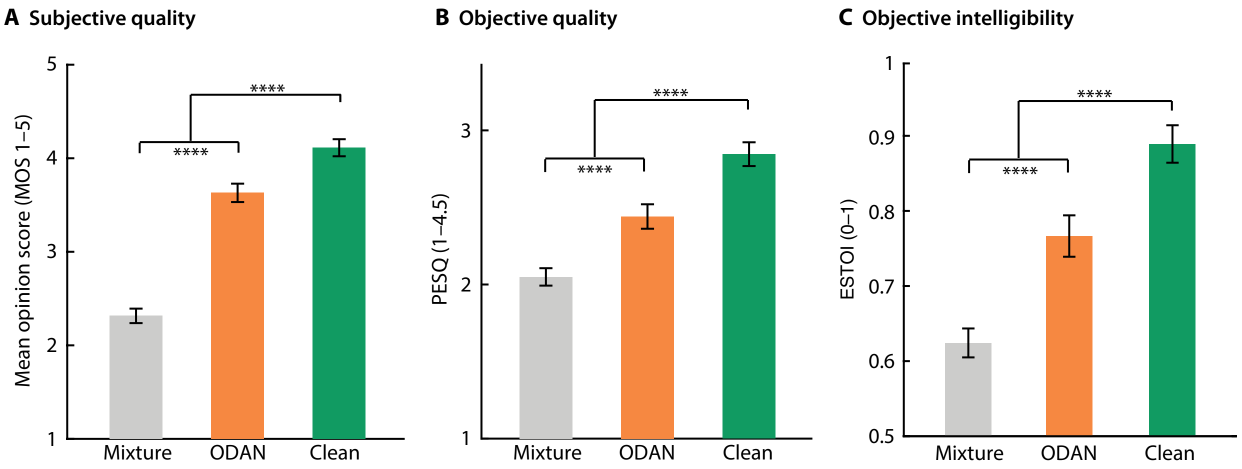


Figure 5.3: **Improved subjective quality and objective quality and intelligibility of the ODAN-AAD system.** (A) Subjective listening test to determine the ease of attending to the target speaker. Twenty healthy subjects were asked to rate the difficulty of attending to the target speaker when listening to (i) the raw mixture, (ii) the ODAN-AAD amplified target speaker, and (iii) the clean-AAD amplified target speaker. The detected target speakers in (ii) and (iii) were amplified by 12 dB relative to the interfering speakers. Subjects were asked to rate the difficulty on a scale of 1 to 5 (MOS). The bar plots show the median MOS ± SE for each condition. The enhancement of the target speaker for the ODAN-AAD and clean-AAD systems was 100 and 118%, respectively ($P < 0.001$). (B and C) Objective quality (PESQ) and intelligibility (ESTOI) improvement of the target speech in the same three conditions as in (A). ★★★★ $P < 0.0001$, *t*-test.

listeners had a stronger preference for the modified audio than for the original mixture. Figure 5.3A also shows a small but significant difference between the average MOS score with the actual clean sources and that with ODAN separated sources (78% versus 56% improvement over the mixture). The MOS values using the clean sources show the upper bound of AAD improvement if the speaker separation algorithm was perfect. Therefore, this analysis illustrates the maximum extra gain that can be achieved by improving the accuracy of the speech separation algorithm (14% over the current system). Figure 5.3B shows a similar analysis when an objective perceptual speech quality measure is used (PESQ) [87], showing a result similar to what we observed in the subjective tests. Together, Fig. 5.3 demonstrates the benefit of using the ODAN-AAD system in improving the perceived quality of the target speaker.

5.4 Discussion

We present a framework for AAD that addresses the lack of access to clean speech sources in real-world applications. Our method uses a real-time, speaker-independent speech separation algorithm that uses deep-learning methods to separate the speakers from a single channel of audio. Then, the separated sources are compared with the reconstructed spectrogram from the auditory cortical responses of the listener to determine and amplify the attended source. The integration of speaker-independent speech separation in the AAD framework enables a brain-controlled hearing assistant system. We tested the system on two unseen speakers and showed improved subjective and objective perception of the attended speaker when using the ODAN-AAD framework.

A major advantage of our system over previous work [160] is the ability to generalize to unseen speakers, which enables a user to communicate more easily with new people. Because ECoG electrodes reflect the summed activity of thousands of neurons in the proximity of the electrodes [168], the spectral tuning resolution of the electrodes is relatively low [169]. As a result, the reconstruction filters that map the neural responses to the stimulus spectrogram do not have to be trained on specific speakers and can generalize to novel speakers, as we have shown previously [12, 165]. Nonetheless, generalization to various noisy, reverberant acoustic conditions is still a challenging problem and requires training on a large amount of data recorded. In addition to increasing the amount of training data and training conditions, separation accuracy can be significantly improved when more than one microphone can be used to record mixed audio. The advantage of enhancing speech with multiple microphones has been demonstrated in Chapter 2, particularly in severely noisy environments or when the number of competing speakers is large (e.g., more than two). In Chapter 6, we will use a binaural speech separation model to significantly enhance the AAD system in more challenging environments.

One major limitation in advanced signal processing approaches for hearing technologies is the limited computation and power resources that are available in wearable devices. Nevertheless, designing specialized hardware that can efficiently implement deep neural network models is an active research area that has recently seen substantial progress [170, 171, 172]. Specialized hardware also significantly reduces the power consumption needed for computation. In addition, hearing aid devices can already perform off-board computation by interfacing with a remote device, such as a mobile phone, which provides another possibility for extending the computational power of these devices [9].

The accuracy of AAD also critically depends on the decoding algorithm being used [173, 174]. For example, the accuracy and speed of decoding can be improved when stochastic models are used to estimate the attention focus using a state-space model [175] instead of the moving average that we used in this chapter. In addition, while we used fixed reconstruction filters derived from the S-T responses, this experimental condition may not always be available. In these scenarios, it is possible to circumvent the need for S-T responses by online estimation of the encoding/decoding coefficients from the responses to the mixture [175, 176], which may lead to more flexible and robust estimation of the decoding filters. Last, decoding methods that factor in the head-related filtering of the sound can also improve the attention decoding accuracy [177]. In Chapter 7, we will take advantage of the advances in self-supervised speech representation learning to improve decoding algorithm.

In summary, our proposed speaker-independent AAD system represents a feasible solution for a major obstacle in creating a brain-controlled hearing device, therefore bringing this technology a step closer to reality. Such a device can help hearing-impaired listeners more easily communicate in crowded environments and reduce the listening effort for normal-hearing subjects, therefore reducing listening fatigue.

Chapter 6: Brain-Controlled Augmented Hearing in Realistic Acoustic Environments

Chapter 5 has introduced a brain-controlled augmented hearing system that decodes the user's brainwaves to selectively amplify the speech of the attended speaker. However, prior auditory attention decoding (AAD) studies have relied on oversimplified scenarios with stationary talkers. In this chapter, we present a realistic AAD task that mirrors the dynamic nature of acoustic settings. This task involves focusing on one of two concurrent conversations, with multiple talkers taking turns and moving continuously in space with background noise. We upgrade the brain-controlled augmented hearing system by incorporating the binaural speech separation model introduced in Section 2.3. Thus, This system is more adept at handling complex acoustic scenarios and can preserve the spatial information of separated speakers. Our subjective and objective evaluations demonstrate that the new brain-controlled augmented hearing system enhances speech intelligibility and facilitates conversation tracking while maintaining spatial cues and voice quality in challenging acoustic environments. This chapter demonstrates the potential of our approach in real-world scenarios and marks a significant step towards developing assistive hearing technologies that adapt to the intricate dynamics of everyday auditory experiences.

6.1 Introduction

In Chapter 5, we have shown that auditory attention decoding can be combined with speech separation techniques to enable brain-controlled augmented hearing technologies. The automatic speech separation algorithm isolates individual talkers from a mixture of talkers in an acoustic scene, while the auditory attention decoding algorithm determines which talker is the attended talker. The attended talker can then be enhanced relative to the background to assist the user of the brain-controlled hearing device.

Past studies have established the feasibility of decoding auditory attention from both invasive [12, 160, 74] and non-invasive [14, 178, 179] neural recordings. Despite these advancements, existing studies predominantly employ overly simplistic acoustic scenarios that do not mimic the real world scenarios [160, 74, 180, 14, 178, 181]. Common experimental setups have been limited to stationary talkers without background noise, and primarily focus on distinguishing between two concurrent talkers. This lack of realism in experimental design is a significant barrier to the generalization of these technologies to everyday life scenarios. Real-world listening involves dynamic conversation involving multiple talkers, often engaged in turn-taking while moving in space, all amidst varying background noises. This chapter aims to bridge this gap by simulating a more realistic experimental paradigm, therefore advancing the field of AAD towards practical applications.

Another important factor that past research has often overlooked is the listeners' desire to track moving talkers in space. This aspect is crucial for natural listening [182] and, thus, for the effectiveness of brain-controlled hearing devices. A successful brain-controlled hearing device must separate speech streams as they move in space while preserving the perceived spatial location of each talker (see Section 2.3). Previous studies of AAD have been based on decoding only the spectrotemporal features of speech. However, recent scientific studies have shown that the human auditory cortex also encodes the location of the attended talker [183, 177] which can potentially lead to the ability to decode the spatial trajectory of attended talkers. This chapter takes a crucial step by investigating whether adding talker trajectories can improve the AAD performance.

Another persistent challenge in AAD research is the difficulty in accurately determining the specific talker to which a subject is attending, especially with high temporal resolution. Previous methods often assume that subjects continuously focus on a pre-designated talker, overlooking the possibility of inadvertent attention shifts [184]. This assumption can lead to mislabeling in data and biasing the performance evaluation of AAD algorithms. This chapter addresses this issue by integrating a behavior measure into our experimental design to ascertain the ongoing focus of the subject more precisely, thereby enhancing the reliability of our data and the validity of our evaluation metrics.

In this chapter, we present a comprehensive and novel approach to AAD that uses complex, dynamic stimuli that more closely resemble real-world acoustic environments. Specifically, we use two concurrent conversations that feature moving talkers and natural background noise, alongside speaker turn-taking among attended and unattended conversations. Furthermore, we introduce a novel task for determining the ground truth labels in attention-focused conversation by requiring the subject to detect deliberately placed repeated words (1-back task) [185, 186]. Lastly, we enhance the brain-controlled augmented hearing system through two key upgrades: (1) We improve the speech separation module by employing the binaural speech separation algorithm described in Section 2.3 which is more robust in complex acoustic scene settings with moving talkers and background noise; (2) We refine the AAD module to utilize both spectral and spatial information for more accurate attention decoding. We show that the proposed system enhances speech intelligibility and facilitates conversation tracking while maintaining spatial cues and voice quality in challenging acoustic environments, hence taking a significant step toward brain-controlled hearing devices in realistic listening environments.

6.2 Materials and Methods

6.2.1 Participants and Neural Recordings

Three new subjects contributed to the data described in this chapter. Subjects 1 and 2 were from North Shore University Hospital (NSUH), and Subject 3 was from Columbia University Irving Medical Center (CUIMC). The participants provided informed consent as per the local Institutional Review Board (IRB) regulations. Subjects 1 and 2 were both implanted with subdural electrocorticography (ECoG) grid and stereo-electroencephalography (sEEG) depth electrodes on their left-brain hemispheres. Subject 3 only had sEEG depth electrodes implanted over their left-brain hemisphere. The procedures followed were identical to those described in Section 5.2.1. Subject 1, 2, 3 had 17, 34 and 42 speech-responsive electrodes respectively as shown in Fig. 6.1.

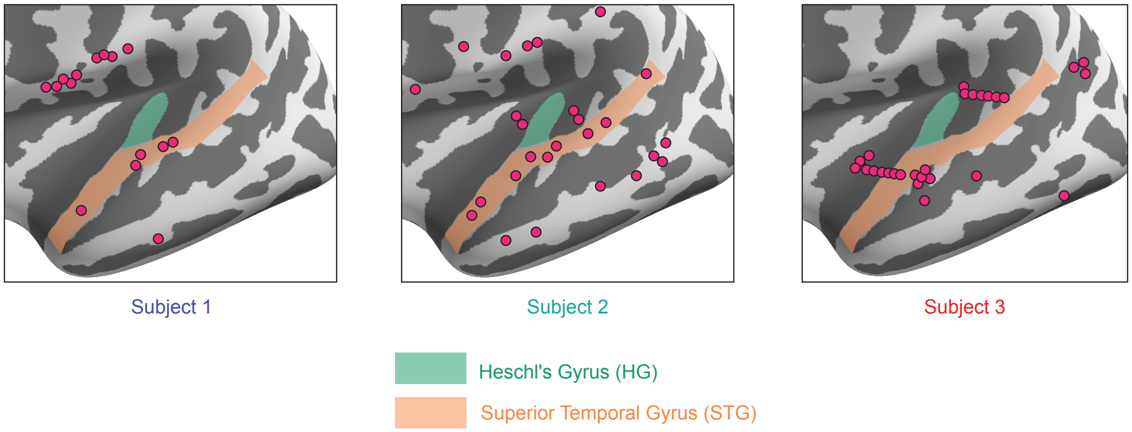


Figure 6.1: Sites of speech-responsive electrodes used for analysis from the three subjects. All subjects had coverage over their left temporal lobe.

6.2.2 Data Preprocessing and Hardware

The neural data of participants from NSUH (Subjects 1 and 2) were recorded using Tucker-Davis Technologies (TDT) hardware using a sampling rate of 1526 Hz. The neural data of the participant from CUIMC (Subject 3) was recorded using Natus Quantum hardware using a sampling rate of 1024 Hz. Neural data was pre-processed following the procedures described in Section 5.2.2.

6.2.3 Stimuli Design and Experimental Paradigm of a Realistic AAD Task

The experiment consisted of 28 multi-talker trials with a mean trial duration of 44.2 s (SD = 2.0 s). The total experiment lasted 26 minutes. As shown in Fig. 6.2a, the trials consisted of two concurrent and independent conversations (one to-be-attended, one to-be-ignored) that were spatially separated and continuously moving in the frontal half of the horizontal plane of the subject. The distances of these conversations from the subject were equal and constant throughout the experiment. Both conversations were of equal power (RMS). Talkers were all native American English speakers. Monaural background noise [187, 188] (either “pedestrian” or “speech babble”) was also mixed along with the conversations at power either 9 dB or 12 dB below the power of

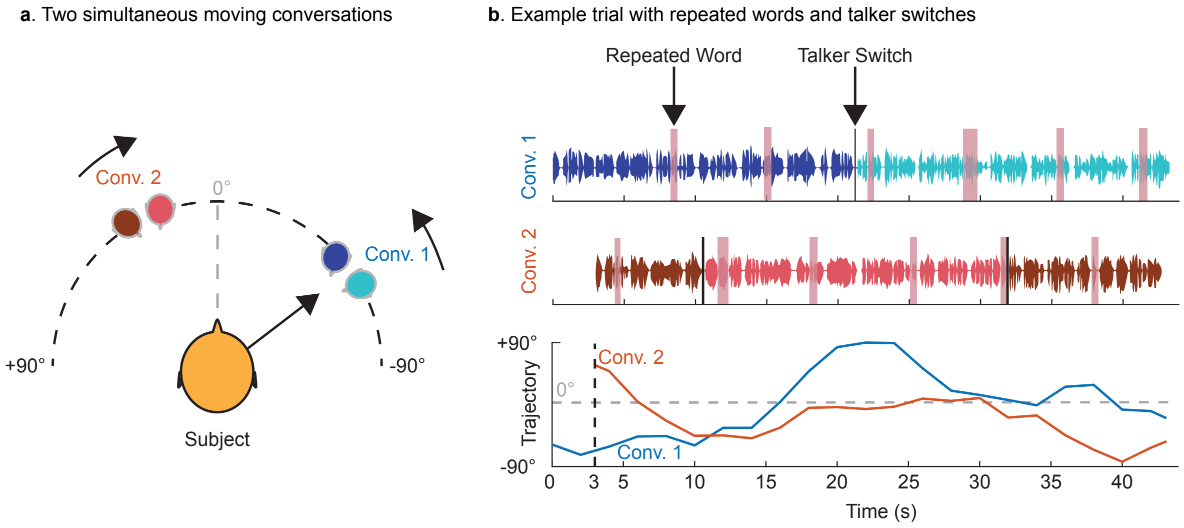


Figure 6.2: **Experiment design**. (a) Every trial consisted of two concurrent conversations moving independently in the front hemifield of the subject. Each conversation had two distinct talkers taking turns. (b) Repeated words were inserted across the two conversations as highlighted in pink. The cued (to-be-attended) conversation had a talker switch at 50% trial time mark whereas the uncued (to-be-unattended) conversation had two talker switches, at 25% and 75% trial time marks.

a conversation stream. Stimuli was delivered to the participants with a sampling rate of 44.1 kHz through stereo earphones (Panasonic RP-HJE120). The to-be-ignored conversation started 3 s later than the to-be-attended conversation. The participants were cued to attend to the conversation that started first.

A total of eight native American English voice actors (four male, four female) were recruited to voice these conversations. These conversations were based on general daily life situations. Every trial consisted of four talkers: two for the to-be-attended conversation (say A and B), two for the to-be-unattended conversation (say C and D). As shown in Fig. 6.2b, in the to-be-attended conversation, a talker switch took place at around 50% trial time mark whereas for the to-be-unattended conversation, two talker switches took place, one at around 25% trial time mark and the other nearly at the 75% trial time mark. Thus, the talker in the to-be-attended conversation would transition from A to B and the talker in the to-be-ignored conversation would transition from C to D to back to C.

In order to check to which conversation a participant might be attending, repeated words were artificially inserted in both the to-be-attended and the to-be-ignored conversations. Participants were asked to press a button upon hearing a repeated word in the to-be-attended conversation. They were expected to ignore the repeated words in the to-be-unattended conversation. The conversation transcripts were force aligned with the audio recordings of the voice actors using the Montreal Forced Aligner tool [189]. The repeated words were inserted in the conversations based on the following criteria: (1) the number of repeated words to be inserted in a conversation of a trial was determined by dividing the trial duration (in seconds) by 7 and rounding the result; (2) for every trial, an equal number of repeated words were inserted in the to-be-attended and the to-be-ignored conversations; (3) a word could be repeated only if its duration was at least 300 ms; (4) to make repeated words sound smooth and natural, a Hanning window of 30 ms was applied to both sides of the audio segment corresponding to the repeated word; (5) the audio segment corresponding to a repeated word was also prefixed and postfixed with 200 ms of silence; (6) the time interval between the onsets of two repeated words in a conversation was constrained to lie between 5.5 s to 9.5 s; (7) there was always one repeated word whose onset was within 1.5 s post talker switch in the to-be-attended conversation. This was done to check if participants tracked the switch in talkers in the to-be-attended conversation; (8) the onset of the first repeated word in a trial was constrained to lie between 5 - 8 s from trial start time. This first repeated word could occur either in the to-be-attended conversation or the to-be-ignored conversation; (9) the minimum time gap between a repeated word onset in the to-be-attended conversation and a repeated word onset in the to-be-ignored conversation was set to be at least 2.5 s. This was done to prevent simultaneous overlap of repeated words in the two conversations and to allow for determining to which conversation a participant was attending to.

Google Resonance Audio software development kit (SDK) was used to spatialize the audio streams of the conversations [190]. The trajectories for these conversations were confined to the frontal half of the horizontal plane of the subject in a semi-circular fashion. In other words, the conversations were made to move on a semi-circular path at a fixed distance from the subject spanning -90 degrees (right) to +90 degrees (left). The trajectories were initially generated with a resolution of 1 degree and a sampling rate of 0.5 Hz using a first order Markov chain. This Markov chain had 181 states (-90 degrees to +90 degrees with a resolution of 1 degree). All states were equally probably of being the initial state. The subsequent samples of a trajectory were generated with a probability transition matrix. The resulting trajectories were smoothed with a moving average of five samples and then stretched to span the whole frontal half plane. The trajectories were further upsampled using linear interpolation to 10 Hz. A pair of trajectories corresponding to a pair of conversations in a trial also followed the following criteria: (1) The spatial separation between the conversations when the second conversation starts was set to be at least 90 degrees; (2) the spatial separation between the conversations during the talker switch in the to-be-attended conversation was ensured to be at least 45 degrees; the correlation of the two trajectories were ensured to be less than 0.5. A total of 1000 trajectory sets (each with 28 pairs, one for each of the 28 trials) were generated based on the above criteria. To have the trajectories span a uniform joint distribution, the set with the highest joint entropy (computed with a bin size of 20 degrees) was chosen as final.

6.2.4 A Brain-Controlled Augmented Hearing System

In this chapter, we propose an innovative brain-controlled hearing system designed for binaural hearing. The system has two microphones to capture the left and right components of the sounds arriving at the ears of the wearer (see Fig. 6.3a). Building upon the foundation laid in Chapter 5, the upgraded framework for the brain-controlled hearing system in this chapter transitions from the monaural speech separation model to a binaural speech separation model, and replaces the stimulus reconstruction-based attention decoder with a canonical correlation analysis (CCA)-based attention decoder (see Fig. 6.3b). The binaural speech separation model, introduced in Section 2.3, as also shown in Fig. 6.3c, separates a binaural mixture of speech streams of two moving talkers (recorded by the binaural microphones) into their individual speech streams while also preserving their spatial cues. As spatial cues are preserved in the separated speech streams of the talkers, the

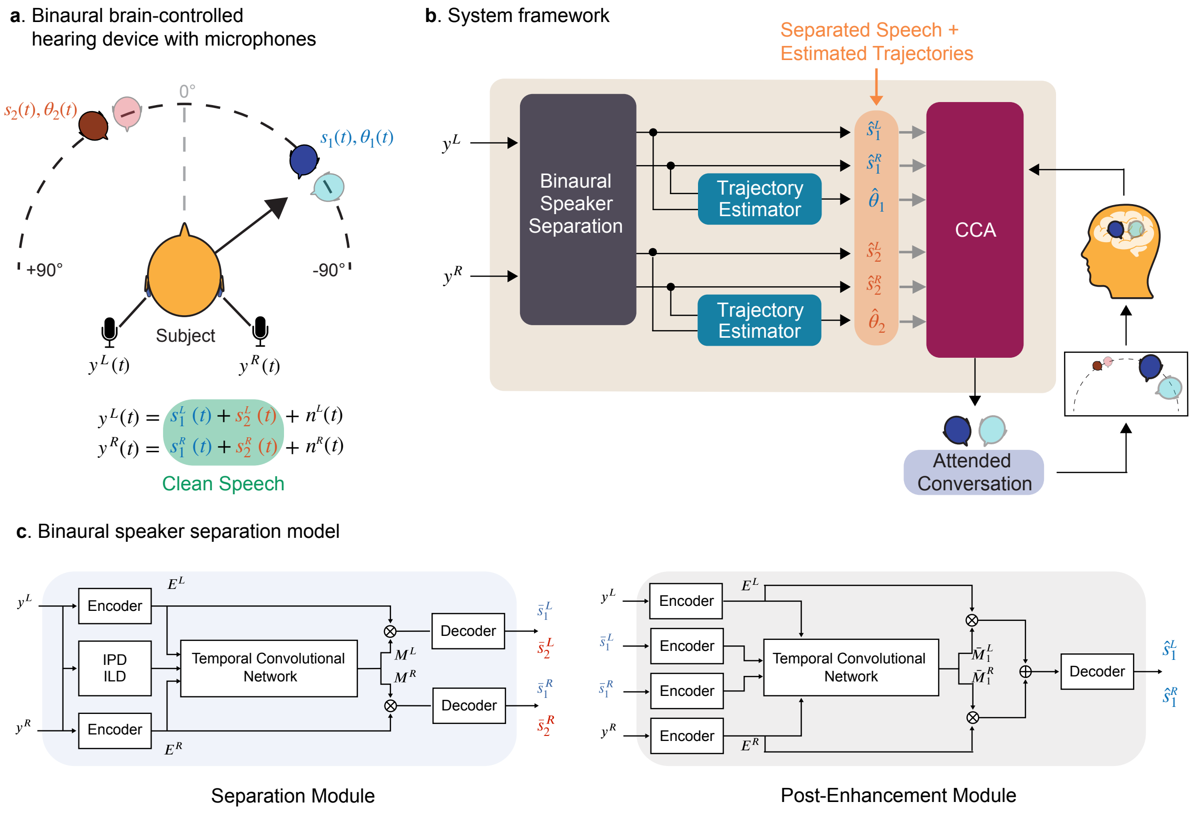


Figure 6.3: **The proposed framework for a binaural brain-controlled hearing device**. (a) The framework requires two microphones, one each on both the left and the right ear. The microphones separately capture the left and the right mixtures of sound sources arriving at the ears. (b) The speaker separation works with these microphone recordings to binaurally separate the speech streams while also estimating the trajectories of the talkers. These outputs are used in combination with the wearer's neural data to decode and enhance the attended talker. (c) The binaural speaker separation model consists of an initial separation module whose outputs are further improved by a post-enhancement module.

model is also able to estimate the trajectories of the moving talkers in the acoustic scene. Auditory attention decoding is enabled by performing CCA which uses the wearer's neural data and the talkers' separated speech and estimated trajectory streams to determine and enhance the attended talker.

In this Chapter, we trained the binaural speech separation model using a different dataset from that in Section 2.3.3. Here, we created 24,000 9.6-second binaural audio mixtures. Each mixture comprised of two moving speakers and one isotropic background noise. The moving speech stimuli were created using the methods described in Section 6.2.3. Speech was randomly sampled from the Librispeech dataset [125]. For half of the training data, we chose pairs of trajectories that spanned uniform distribution (quantified by joint entropy); and for another half of the training data, we chose pairs of trajectories whose average distance difference was smaller than 15 degrees to enhance the separation model's ability to handle closely spaced moving speakers. We randomly chose noise from DEMAND dataset [98]. The SNR, defined as the ratio of the speech mixture in the left channel to the noise, ranged from -2.5 to 15 dB. All sounds were resampled to 16 kHz.

6.2.5 Canonical Correlation Analysis

In Chapter 5, we used the stimulus reconstruction method for attentional decoding where only the spectrogram of the attended talker is estimated. In this chapter, we used CCA [174] to predict the attended talker. From the stimuli side, the inputs involved both talker spectrograms and trajectories. We chose a 20-bin mel spectrogram representation obtained with a window duration of 30 ms and a hop size of 10 ms. Audio was downsampled to 16 kHz before mel spectrogram extraction. The mel spectrograms of left and right channels were concatenated along the bin dimension. All trajectories were upsampled to 100 Hz from 10 Hz to match the sampling rate of the neural data. Trajectories were pooled across all trials and normalized. Spectrograms were also normalized on a bin-by-bin basis. We chose a receptive field size of 500 ms for neural data and 200 ms for stimuli spectrograms and trajectories. The starting sample of these receptive fields were aligned in time. Time-lagged matrices were then generated individually for neural data, trajectory and spectrograms.

As done in a previous study [174], principal component analysis (PCA) was applied individually to time-lagged versions of both spectrogram and trajectory. PCA was also applied to the time-lagged neural data matrix. The top PCA components explaining at least 95% of the variance were retained. This was done to reduce the risk of overfitting in CCA.

During training, the CCA models simultaneously learn forward filters on attended talker's clean speech spectrogram and trajectory (after PCA) and backward filters on the neural data (after PCA) such that upon projection with these filters, the neural data and the attended talker stimuli would be maximally correlated. During testing, these learnt filters are applied to the neural data (after PCA) as well as to every talker's speech spectrogram and trajectory (after PCA). The talker which yields the highest correlation score (based on voting of the top three canonical correlations) was determined as the attended talker.

We used behavioral measurements to correct the “attended” and “unattended" labels for the two conversation streams . For trials in which two or more repeated words were detected in the uncued conversation, the corresponding portions (bounded by button press timings) of the cued to-be-attended and uncued to-be-unattended stimuli were swapped before model training and evaluation. For models trained without correction, no such swapping was done based on behavior.

6.2.6 Psychoacoustic Experiment

The online psychoacoustic experiment to evaluate the performance of the brain-controlled hearing system was conducted with 24 self-reported normal hearing participants from Amazon MTurk. These participants were native speakers of American English located in the US. The experiment lasted for a total of 30 minutes per participant and each participant was paid with 10 dollars. All participants were required to wear stereo earphones.

Every MTurk participant listened to a total of 15 trials, 5 trials from each of the following conditions.

1. System Off: The raw mixture stimuli that was played to the subjects from whom neural data was recorded.

2. System On (Separated): Mixture in which the attended talker, as determined by the neural signatures, was enhanced using the output of the binaural speaker-separation model.

3. System On (Clean): Mixture in which the attended talker was enhanced using clean ground truth speech.

The trial order was randomized and the participants were unaware of the conditions assigned to the trials. Enhanced mixtures were generated by suppressing the un-attended talker and the background noise in the mixture by the same scale factor such that the resulting power difference between the attended and the unattended talker was 9 dB. Like the iEEG participants, the MTurk participants were also instructed to follow the cued conversation (conversation that starts first) and press space bar on their keyboards upon hearing the repeated words in the conversation being followed. After every trial, the participants were prompted with the following four questions:

1. Comprehension: A multiple choice question based on the content in the to-be-attended conversation with a single correct answer. This tested the intelligibility of the conversations.

2. Difficulty: Participants were asked to rate how difficult or easy it was for them to follow the cued conversation on a scale from 1 to 5 (1 = very difficult, 2 = difficult, 3 = neutral, 4 = easy, 5 = very easy).

3. Sound Localization: The last three seconds of the trial was allowed to be replayed multiple times by the participants. Participants were asked to indicate from one of five equally partitioned sectors of the frontal half plane (left, front left, center, front right, right) where the cued conversation ended. This tested weather the spatial information of the separated speakers were accurately preserved.

4. Voice Quality: Participants were also asked to rate the quality of voices in the cued conversation on a scale from 1 to 5 (1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent).

6.3 Results

6.3.1 Behavioral Data Analysis

The push button responses of subjects to repeated words in the conversation being followed help in determining to which conversation a subject was attending. A repeated word in a conversation was considered as correctly detected only if a button press was captured within two seconds of its onset. As shown in Fig. 6.4, all subjects tracked more than 65% of the repeated words in the cued (to-be-attended) conversation. We assign these as hits. However, we see that subjects also

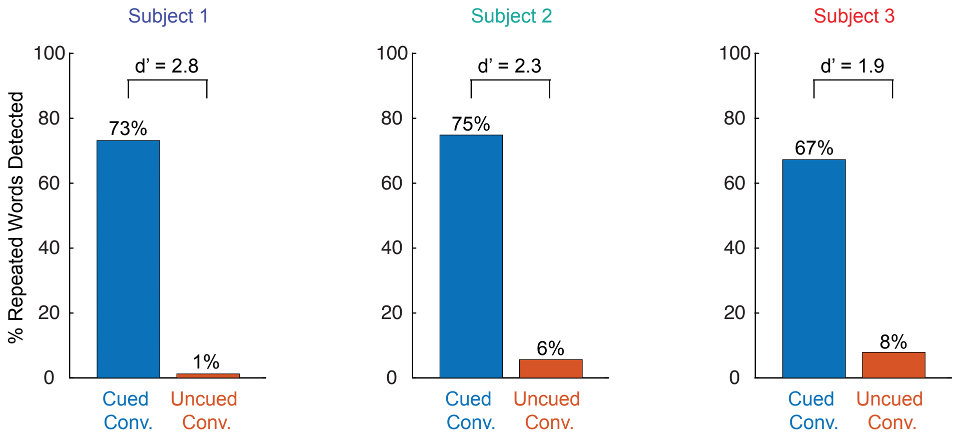


Figure 6.4: Proportion of repeated words detected in the cued (to-be-attended) and uncued (to-be-unattended) conversations by the subjects across all trials. Subjects mostly attend to the cued conversation, but sometimes they also pay attention to the uncued one.

tracked a non-zero fraction of repeated words in the uncued (to-be-unattended) conversation (false alarms) indicating that there might have been occasions when the subjects were attending to the uncued (to-be-unattended) conversation. We combined the hit rate and false alarm rate for each subject to generate a sensitivity index (d') inspired by signal detection theory [185, 191] (SDT). Sensitivity index for each subject was calculated as: d' = z(False Alarm Rate) – z(Hit Rate), where z(x) is the z-score corresponding to the right-tail p-value of x [191]. Subjects were ranked based on their sensitivity indices (S1: 2.8, S2: 2.3, S3: 1.9).

6.3.2 Auditory Attention Decoding

Section 6.2.5 introduces the CCA algorithm that compares neural signals with both speech spectrogram and trajectory of each talker to predict the attended talker. Subject-wise CCA models were trained, and their performance was evaluated using leave-one-trial-out cross validation, i.e., training on N - 1 trials and testing on the windows from the N-th trial. We evaluated auditory attention decoding accuracies for all subjects for a range of window sizes from 0.5 s to 32 s for the following two stimuli versions:

1. Clean Stimuli: Using the clean (before mixing) ground truth speech spectrograms and trajectories of individual talkers in the acoustic scene.

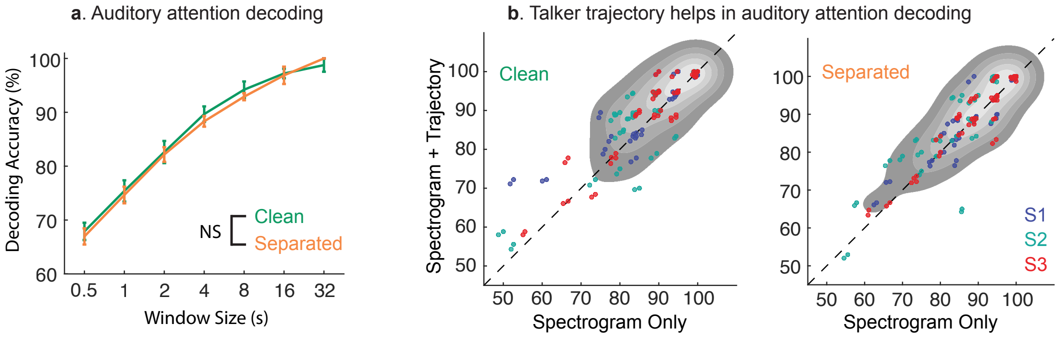


Figure 6.5: **Evaluating AAD performance.** (a) AAD accuracies averaged across subjects as a function of window size. The decoding accuracies are comparable between the clean and separated versions (Wilcoxon signed-rank test, p-val = 0.13). Error bars indicate the standard error of mean. (b) Scatter plots comparing trial-wise AAD accuracies for a window size of 4 s when using only spectrogram vs spectrogram + trajectory. Each point represents a trial. AAD accuracies improved significantly when talker trajectories were also incorporated in addition to their speech spectrograms for both clean (paired *t*-test, p-val = 0.002) and separated (paired *t*-test, p-val = 0.010) versions.

2. Automatically Separated Stimuli: Using the speech spectrograms and estimated trajectories of talkers yielded by the binaural speech separation model.

Figure 6.5a shows the attended talker decoding accuracies averaged across subjects as a function of window size for both clean and separated versions after correcting for behavior. For both versions, the attended talker decoding accuracies increase as a function of window size. This is expected since with larger window sizes, more information is available to determine the attended talker. Stimuli version had a very small effect on the AAD accuracies across subjects and window sizes (Wilcoxon signed-rank test, p-val = 0.13). This indicates that the AAD performance with automatically separated stimuli is as good as the performance with original clean stimuli (in Fig. 6.5a), confirming the efficacy of the proposed speech separation module.

We studied the improvement in AAD performance when talker trajectories are included in addition to talker spectrograms. For this comparison, we trained and tested CCA models (post behavior correction) with only talker spectrograms without trajectories. As shown in Fig. 6.5b, we found that trial-wise AAD performance improved when talker trajectories were also incorporated in addition to talker spectrograms for both clean (paired *t*-test, p-val = 0.002) and automatically

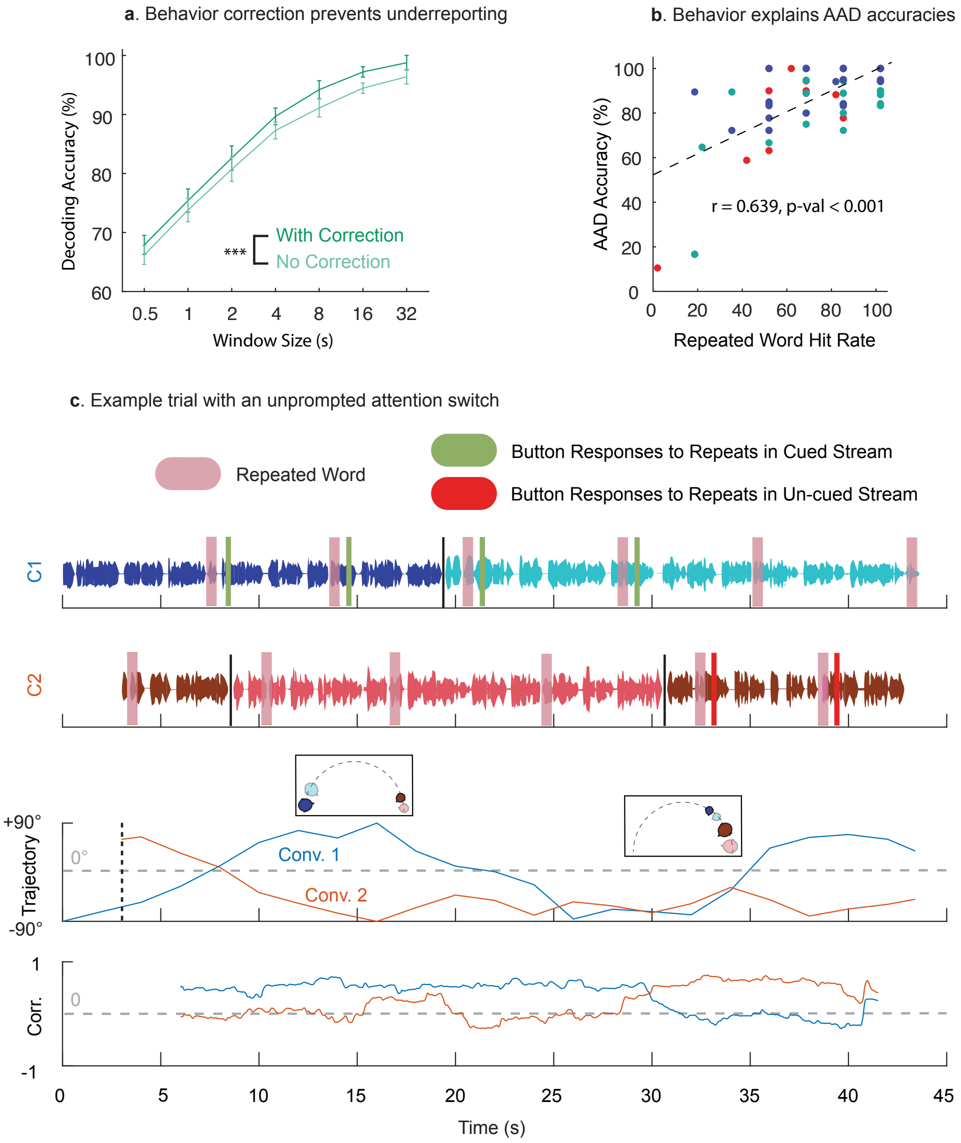


Figure 6.6: **Behavior measurement tracks the actual attention of the listener.** (a) A separate set of models were trained without correcting for behavior. The decoding accuracies are plotted for the clean version of speech for both with and without behavior correction. Not correcting for behavior can lead to significant underreporting of AAD performance (Wilcoxon signed-rank test, p-val < 0.001) (b) For models trained without correcting for behavior, trial-wise behavioral performance and AAD accuracies are significantly correlated (Pearson's r = 0.639, p-val < 0.001). (c) An example trial from one of the subjects who shifts attention from the cued conversation (Conv. 1) to the uncued conversation (Conv. 2) in the middle of the trial. Repeated words in the conversation streams are shaded in pink. Button press responses to the repeated words are shown in green (red) for the cued (uncued) conversation. The last plot shows the first canonical correlation for both the conversation streams obtained by continuously sliding a 4 s window. Behavior is well correlated with the canonical correlations.

separated (paired *t*-test, p-val = 0.010) versions of the stimuli.

Lack of having a behavioral measure and not correcting for the same can lead to underreporting of AAD performance. To study this, we also trained a set of CCA models assuming that the subjects always paid attention to the cued (to-be-attended) conversation. Figure 6.6a compares the AAD performance for clean stimuli when correcting and not correcting for behavior. Not correcting for behavior significantly hurts AAD performance (Wilcoxon signed-rank test, p-val < 0.001). This is also true when evaluating with the automatically separated version of the stimuli (Wilcoxon signed-rank test, p-val < 0.001).

Next, for the models trained without correcting for behavior, we examined whether the behavioral performance on the repeated word detection task could explain the AAD performance on a trial-by-trial basis. We first computed the proportion of repeated words detected in the cued conversation (hit rate) for each trial and for each subject. We also computed corresponding trial-wise AAD accuracies for a window size of 4 s. As shown in Fig. 6.6b, we found that hit rate on the repeated word detection task was significantly correlated with the trial-wise AAD accuracies (Pearson's r = 0.639, p-val < 0.001). Figure 6.6c shows an example trial from one of the subjects who, based on behavioral responses, was initially attending to the cued (to-be-attended) conversation and then later attends to the uncued (to-be-unattended) conversation after the conversations cross in space. The canonical correlations mapping the neural data with both the cued and uncued stimuli also capture this shift of attention from one conversation to the other. Thus, the repeated word detection task helps explain AAD performance on a trial-by-trial basis.

6.3.3 System Dynamics During Talker Transitions

Turn-takings during conversations create talker switches in the attended conversation. For good user experience, it is important that the system tracks the talker switch and seamlessly enhances the new attended talker. Our experiment paradigm, inspired by real-world settings, had asynchronous talker switches in both to-be-attended and to-be-unattended conversations (see Section 6.2.3). Figure 6.7a shows the system was able to seamlessly track turn-takings in conversations.

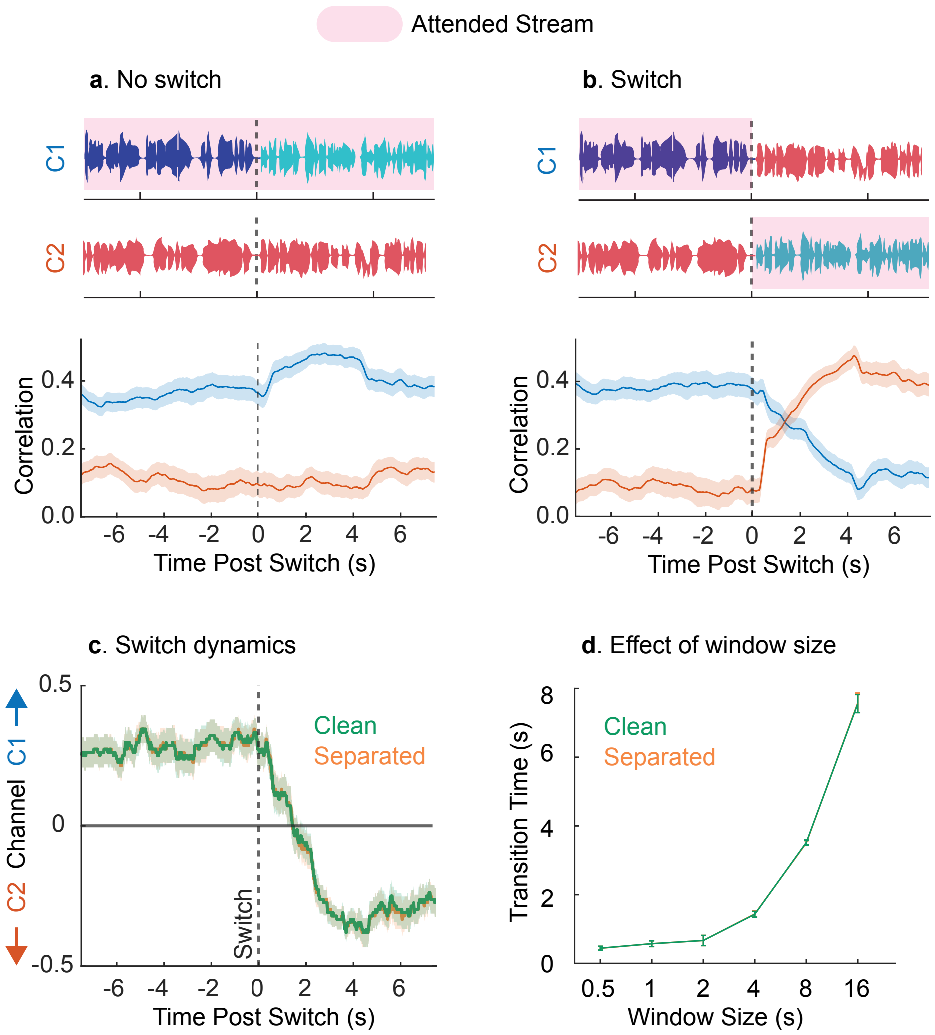


Figure 6.7: **System dynamics during talker transitions.** (a) The proposed system seamlessly tracks turn-takings. This is facilitated by the speaker separation module which places talkers in a conversation on the same output channel by relying on location and talker continuity cues. Attended conversation is highlighted with a pink shade. Correlations showed are the average of the top three canonical correlations for separated version of the stimuli. (b) Attention switch from one conversation to another can be simulated by swapping the output channels of the binaural separation system. (c) Channel preference dynamics after simulated attention switch for a decoding window size of 4 s. (d) Transition times as a function of decoding window size. No significant differences were observed between the clean and separated versions (Wilcoxon signed-rank test, p-val = 0.70). Error bars in all plots indicate the standard error of mean.

In some cases, the wearer of the hearing device might switch attention from a conversation at a particular location to another conversation at a different location. To study how our system responds in such cases, we artificially swapped the outputs of the binaural speech separation system at the point of talker switch in the cued conversation, as shown in Fig. 6.7b. Since we combine the results of the top three canonical correlations based on voting to determine the attended talker or channel, we define a metric channel preference index (CPI), i.e.,

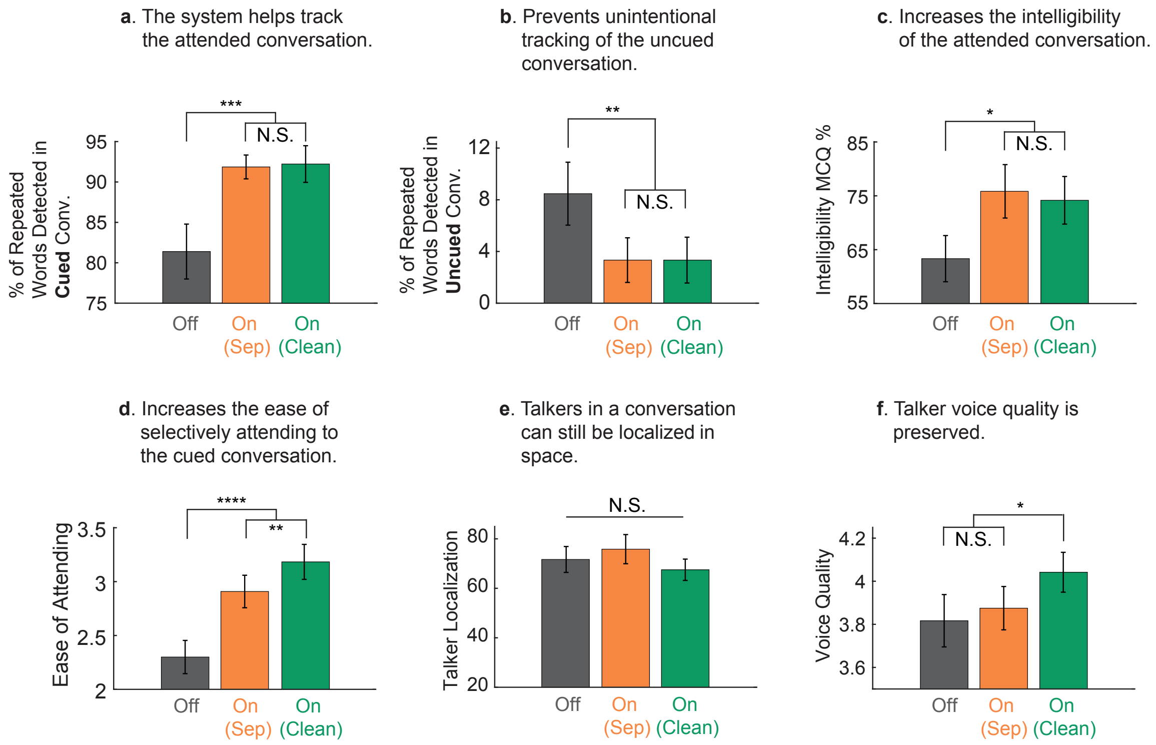
$$\text{CPI} = \frac{\text{\# of votes favoring Channel 1}}{3} - 0.5. \tag{6.1}$$

Thus, a positive CPI would indicate a preference to Channel 1 whereas a negative CPI would indicate a preference to Channel 2. In Fig. 6.7c, we show the CPI averaged across trials for one of the subjects (S3) when attention switch is simulated. We define the transition time as the time point where the average CPI crosses 0. Figure 6.7d shows the transition times (averaged across subjects) as a function of window size for both clean and separated versions. No significant difference was found in the transition times across subjects and window sizes between the clean and separated versions (Wilcoxon signed-rank test, p-val = 0.70).

6.3.4 Evaluation of System Performance

Part A: Subjective

We recruited human participants to evaluate the performance of the proposed system (see Section 6.2.6). As shown in Fig. 6.8a, under both "system on" conditions, the repeated word detection accuracy in the cued conversation is enhanced when compared to the "system off" condition (paired t-test, p-val < 0.001), whereas for the uncued conversation (in Fig. 6.8b), the detection accuracy is reduced (paired t-test, p-val < 0.01). This means that the system helps track the cued conversation and prevents unintentional tracking of the uncued conversation. We also find that intelligibility of the cued conversation is significantly enhanced under the "system on" conditions (in Fig. 6.8c, paired t-test, p-val < 0.05). No significant differences are observed between the clean and separated versions of the "system on" condition. Ease of attending to the cued conversation increases from "system off" condition to "system on with separated speech" condition (paired t-test, p-val < 0.0001) to "system on with clean speech" condition (paired t-test, p-val < 0.01), as shown in



a. The system helps track the attended conversation.

b. Prevents unintentional tracking of the uncued conversation.

c. Increases the intelligibility of the attended conversation.

d. Increases the ease of selectively attending to the cued conversation.

e. Talkers in a conversation can still be localized in space.

f. Talker voice quality is preserved.

Figure 6.8: **Subjective evaluation of system outputs show enhanced tracking of the cued conversation, improved intelligibility and retention of talker cues and voice quality.** (a) Repeated word detection accuracy in the cued conversation increases significantly when the system is turned on for both clean as well as separated versions (paired t-test, p-val < 0.001). (b) Repeated word detection accuracy for the uncued conversation drops significantly when the system is turned on (paired t-test, p-val < 0.01). (c) Intelligibility of the cued conversation is significantly increased under the system on conditions (paired t-test, p-val < 0.05) (d) Attending to the cued conversation is easier under the system on conditions (paired t-test, p-val < 0.0001). (e) Participants can localize talkers in space equally well in all conditions. (f) No significant difference in voice quality ratings was observed between the system off condition vs the system on with separated speech condition. However, participants rated the voice quality of the system on with clean speech condition to be relatively higher (paired t-test, p-val < 0.05). Error bars in all plots indicate the standard error of mean.

Fig. 6.8d. Surprisingly, no differences in voice quality of the talkers in the cued conversation were observed between the “system off” and the “system on with separated speech” condition (in Fig. 6.8f). However, participants rated the voice quality in the “system off with clean speech” condition higher than the other two (paired t-test, p-val < 0.05). These results indicate that a scope for improvement exists for the binaural speech separation model and its upper bounds (when there is ideal separation) are captured by the “system on with clean speech” condition. The ability to

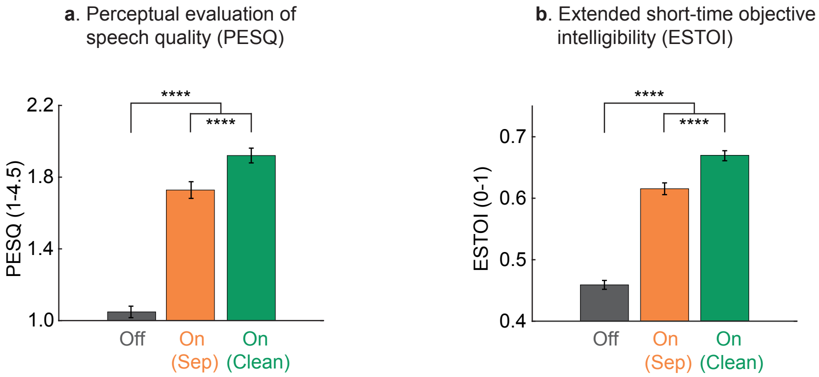


Figure 6.9: **Improved objective quality and intelligibility.** (a and b) Both PESQ and ESTOI scores increase from "system off" condition to "system on with separated speech" condition to "system on with clean speech" condition (paired t-tests, p-val < 0.0001). Error bars in all plots indicate the standard error of mean.

localize talkers in space, as shown in Fig. 6.8e, was comparable across all the three conditions highlighting retention of the attended talker spatial cues when the system is turned on. In summary, the system helps follow the conversation of interest, increases its intelligibility and the ease of attending to it while also preserving spatial cues.

Part B: Objective

In addition to subjective evaluation, we also performed an objective evaluation where the same system simulated outputs in the subjective evaluation were compared with their corresponding clean to-be-attended conversation waveforms (as reference) to calculate narrowband MOS-mapped Perceptual Evaluation of Speech Quality (PESQ) [87] and Extended Short-Time Objective Intelligibility (ESTOI) [88] scores. As expected, in Fig. 6.9, we see a significant improvement in these scores as we progress from "system off" condition to "system on with separated speech" condition to "system on with clean speech" condition (paired t-tests, p-val < 0.0001).

6.4 Discussion

We introduced a novel AAD experimental paradigm that diverges from existing studies by incorporating concurrent conversations with natural turn-takings where talkers move in space amidst background noise. This approach represents a substantial advancement in creating realistic auditory scenarios for AAD research. Our binaural speaker separation model successfully separated these dynamic conversations into individual streams while preserving talker spatial cues and suppressing background noise. Additionally, the speech separation system provides real-time talker trajectory to the AAD algorithm, enhancing its decoding accuracy and speed. The use of repeated word detection tasks across the conversations provided a robust ground truth label for the attended conversation with a high temporal resolution and explained AAD performance on a trial-by-trial basis. Evaluations of the proposed system revealed improved tracking of the attended conversation and increased its intelligibility while preserving the perceived location of each talker in space.

The primary aim of this chapter was to address the limitations of previous AAD research that predominantly assumed two stationary talkers [160, 180, 14, 179], thereby restricting the applicability of such research to real-world scenarios. In realistic acoustic scenes, we normally listen to simultaneous conversations which can involve multiple talkers. Our research extends previous work by replacing concurrent talkers with concurrent conversations involving natural turn-taking. By utilizing the speaker-independent speech separation model introduced in Section 2.3 that leverages both spatial and spectro-temporal information, our research marks a significant step toward creating an immersive listening experience that closely mimics natural environments. This separation model not only separates the speech of moving talkers but also allows listeners to accurately track their locations, an aspect crucial for realistic AAD applications. An essential contribution of our study is that incorporating real-time talker trajectories estimated by speech separation algorithm in addition to spectrotemporal information can improve AAD accuracy and speed [183, 192, 193, 194]. Further research is needed to distinguish listener motion-induced from talker motion-induced acoustic change and how it could be encoded differently in the human auditory cortex.

Another contribution of this chapter is introducing a behavioral task of repeated word detection across conversations, allowing us to identify the actual attended conversation with high temporal resolution. This method addresses a common issue in previous AAD studies where subjects' attention could inadvertently shift to the unattended stream [184], leading to mislabeled data and affecting the training and evaluation of AAD models. By incorporating a behavioral measure into our experiment design, we have enhanced the accuracy of determining the attended talker or conversation. In future AAD studies with moving talkers, a higher degree of temporal resolution can be achieved by asking the subjects also to report the spatial trajectory of the conversation followed. Additionally, further research is needed to investigate the difference between endogenous and exogenous auditory attention switches and how they may be decoded differently [195].

While our study focused on neural activity in the high gamma band, incorporating low-frequency neural activity, which has been shown to track motion and attention, could improve AAD accuracies. Prior invasive [162] and non-invasive [14, 13] AAD studies have shown signatures of auditory attention (via tracking of the envelope of the attended speech) in the lower frequencies (1 – 7 Hz). A recent study [196] also showed that low-frequency neural activity also tracks the location of the attended talker, especially in delta (< 2 Hz) phase and alpha (8-12 Hz) power. Including low-frequency neural signals might provide a more comprehensive understanding of the neural underpinning of auditory attention and enhance the performance of AAD systems.

A critical aspect of future research should involve transitioning to a real-time, closed-loop system. This requires the integration of speech separation and AAD components to work synchronously in a causal, real-time manner. Furthermore, determining how to optimally manipulate the acoustic scene based on the decoded attended talker remains an area for further investigation. Such acoustic modifications should help the listener follow the attended conversation while still maintaining the ability to switch to the unattended one. Our experiment design could be further aligned with real-world scenarios by introducing more complex motion patterns for talkers, such as radial motion and motion pauses. This would add a layer of complexity to the auditory scene, presenting conversations with time-varying power and potentially challenging the current speaker separation model. Addressing this challenge may involve retraining or fine-tuning the model on datasets with these characteristics.

A brain-controlled hearing device that can quickly and accurately adapt to changes in the listener's attention is a challenge that may be more effectively addressed with invasive neural recording techniques. However, a critique of our approach is the reliance on invasive neural recordings which might be perceived as less accessible. Considering the rapid advancements in speech BCI research involving invasive neural recordings [72, 73], these methods are becoming increasingly common and feasible. The precision and speed offered by invasive recordings are currently unmatched by non-invasive techniques, making them essential for exploring the upper limits of AAD performance. While future research continues to explore less invasive or alternative neural recording methods, our current focus on invasive recordings is crucial for advancing the field and setting benchmarks for performance of these systems and establishing minimum required performance for listeners to prefer AAD functionality.

This chapter contributes significantly to AAD research and brain-controlled hearing devices by introducing more realistic experimental paradigms and advancing the technology toward practical applications. The insights from this research enhance our understanding of auditory attention in complex environments and pave the way for future innovations in assistive hearing technologies.

Chapter 7: Improved Decoding of Attentional Selection in Multi-Talker Environments with Self-Supervised Learned Speech Representation

In Chapter 5 and 6, we have used auditory attention decoding (AAD) technique to identify the talker that a listener is focused on in a noisy environment. This is done by comparing the listener's brainwaves to a representation of all the sound sources to find the closest match. The representation is typically the spectrogram of the sounds. However, it is uncertain whether this representation for AAD is optimal. In this chapter, we examine the use of self-supervised learned speech representation in improving the accuracy and speed of AAD. We used WavLM to extract a latent representation of each talker and trained a spatiotemporal filter to map brain activity to intermediate representations of speech. During the evaluation, the reconstructed representation is compared to each speaker's representation to determine the target speaker. Our results indicate that speech representation from WavLM provides better decoding accuracy and speed than the speech envelope and spectrogram. Our findings demonstrate the advantages of self-supervised learned speech representation for auditory attention decoding and pave the way for developing better brain-controlled hearable technologies.

7.1 Introduction

Auditory attention decoding (AAD) uses brain activity to predict which talker the listener is attending to. Most AAD algorithms reconstruct a representation of speech from the brain activity and compare it to all the talkers in the environment [14, 12]. The talker with the highest correlation is considered the attended talker. Typically, the representation of sound that is used in AAD is either the envelope or spectrogram. However, it is not clear if either of these are optimal for neural decoding. A good representation of sound for AAD should be easily reconstructed from brain activity. In particular, learning a complex nonlinear mapping from brain signals to the representation can be challenging due to the limitations in recording the brain signals. Additionally, a good representation should have a stronger correlation with the attended speaker than with unattended speakers.

Self-supervised representation learning (SSL) for speech has been successfully applied in many applications [197, 198, 199]. SSL learns representations through designed pretext tasks, where the input and learning targets are derived from the input signal itself. Because of this, SSL can be easily scaled up with a large amount of unlabeled speech data. The self-supervised learned representations are often used as input features for downstream tasks to reduce the need for a large amount of labeled training data and improve task performance. Studies have shown that the learned speech representations can improve various downstream tasks such as speech recognition, speaker identification, and intent classification [200].

Several self-supervised speech representation learning approaches have been proposed recently, with wav2vec 2.0 [197] and HuBERT [198] being two of the most well-known. Both models have a similar architecture but differ in their pre-training strategies. Wav2vec 2.0 uses a contrastive loss to differentiate between positive and negative samples, while HuBERT uses an offline clustering approach to assign labels to speech units, and then trains the model through a BERT-like masked speech prediction task. This forces the model to learn both acoustic and language features from unlabeled speech data. WavLM [199], a variant of HuBERT, adds a speech denoising task during pre-training to improve its ability to handle non-ASR tasks such as speech diarization and separation. WavLM Large, trained on 94k hours of diverse speech data, outperforms previous self-supervised speech models on SUPERB [200], demonstrating its high capacity to model speech speaker, content, and semantics. Given its strong performance, we selected WavLM Large for this study, as it has the best potential to improve attended-or-unattended talker classification accuracy. Decoding attentional selection is therefore regarded as a downstream task of WavLM.

A recent study indicates that the functional hierarchy of latent layers of a self-supervised speech model aligns well with the cortical hierarchy of speech processing [201]. Additionally, the learned representations have been found to be more effective at predicting cortical responses to speech [202] and being predicted from them [203], compared to hand-engineered acoustic features, further motivating the idea that they may be related to attention and superior to traditional acoustic features used in the AAD task.

In this chapter, we propose to use self-supervised learned speech representations to improve the neural decoding of attentional selection. We used an intermediate layer of the pre-trained WavLM model as the reconstruction target from the brain signals instead of using traditional speech envelope and spectrogram features. Our experimental results show a significant improvement in decoding accuracy when using these learned speech representations. Furthermore, we adapted the WavLM model to a causal setup for real-time implementation testing and showed that it still surpasses speech envelope and spectrogram features, suggesting that transformer-based self-supervised representations excel as candidates for brain-controlled hearable devices.

7.2 Material and Method

7.2.1 Neural Data Acquisition and Preprocessing

We used the same materials as in Chapter 6. For details on participants and neural recordings, refer to Section 6.2.1; for data preprocessing information, see Section 6.2.2; and for specifics on stimuli design, see Section 6.2.3.

7.2.2 Extraction of Speech Representations

The WavLM Large model [1] is composed of a convolutional encoder and 24 transformer layers. The convolutional encoder converts a waveform sampled at 16 kHz to a feature sequence at a 50 Hz framerate (one frame every 20ms), with each frame encoding about 25 ms of the waveform. Each transformer layer has an embedding dimension of 1024 and 12 self-attention heads.

We used WavLM to extract the latent representation $X \in \mathbb{R}^{1024*T}$ of speech waveforms from

[1] The pre-trained model can be found at `https://github.com/microsoft/unilm/tree/master/wavlm`

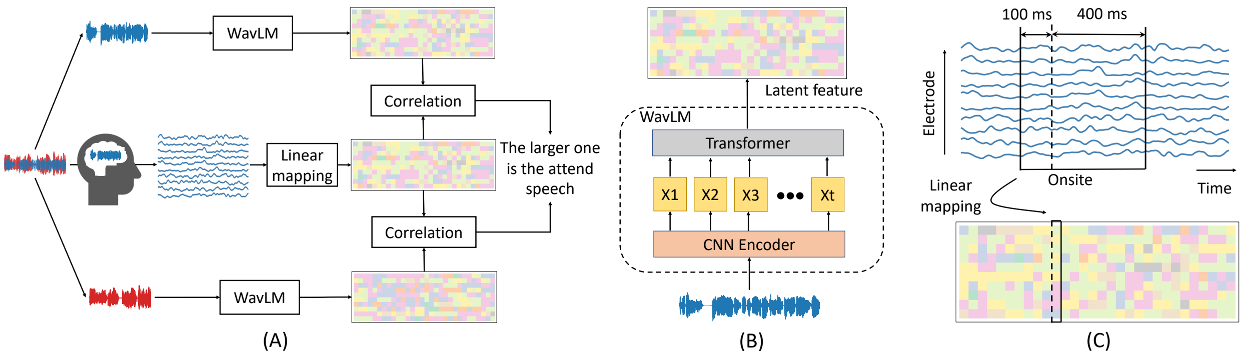


Figure 7.1: **Diagram showing the process of auditory attention decoding.** (A) Neural activity is monitored while a subject listens to a mixture of two talkers and focuses on one of them. WavLM extracts the representation of individual talkers obtained through a speech separation model. The predicted representation from the neural activity is compared to the representation of individual talkers to determine the most similar talker. (B) WavLM consists of a CNN encoder and transformer, producing layers of features. One intermediate layer is used as the speech representation. (C) Linear spatiotemporal filters map the neural activity, with time-lags ranging from −400 ms to 100 ms, to the learned representation.

the i-th layer, where T is the number of time frames. We upsampled X to 100 Hz to match the rate of the neural data. Since our speech duration is long, we limited the attention span of each frame in the transformer layers to 6 seconds ($\sim$ 300 time frames) with 3 seconds before and 3 seconds after the frame. However, WavLM is a noncausal model because each time frame attends to future time frames. For a fair comparison with the speech envelope and spectrogram, and to enable real-time AAD, we modified WavLM for a causal configuration. First, we set the attention weights of all the future frames as zero to force each frame to only attend to the past 6 seconds, and we refer to this model as “WavLM w/ causal ATT". The transformer layer in WavLM is equipped with a convolution-based position embedding where the convolution operation has access to the future frames, which results in noncausal computation. To avoid this, we changed the noncausal convolution to a causal convolution, and the resulting model is referred to as “WavLM w/ causal ATT & PE". Note that we did not finetune WavLM after we modified the attention weights or positional embedding. In case of random effects that cause performance gain for WavLM, we added WavLM with random initialization as a control.

7.2.3 Representation Reconstruction for Decoding Attention

We employed the linear reconstruction method as introduced in Section 5.2.5, with a slight modification: we reconstructed representations of both attended and unattended talkers. This adjustment is based on previous research [13, 14] which demonstrated the feasibility of separately extracting attended and unattended speech from neural data. Specifically, two subject-wise liner spatiotemporal filters G_A and G_u were learned to map neural activity R to the speech representations of the attended ($\hat{X}_A$) and unattended ($\hat{X}_U$) talkers,

$$\hat{X}_A(n,t) = \sum_e \sum_\tau G_A(n,e,\tau) R(e,t-\tau) \tag{7.1}$$

$$\hat{X}_U(n,t) = \sum_e \sum_\tau G_U(n,e,\tau) R(e,t-\tau), \tag{7.2}$$

where n is the channel index of the representation, e is the neural electrode index, and τ is the time lag, ranging from -400 ms to 100 ms in this study. The linear filters were optimized by minimizing the mean-squared errors between the reconstructed and the actual representations.

A leave-one-out cross-validation approach was used, wherein the subject-wise filters were trained on N - 1 trials and used to reconstruct representations $\hat{X}_A, \hat{X}_U$ on the left out trial. We calculated Pearson's correlation coefficient between the reconstructed representations $\hat{X}_A, \hat{X}_U$ and the representations of two talkers X_{sp1}, X_{sp2}. The correlation coefficient is estimated across a window of seconds, which is referred to as the decoding window duration. We used sliding window of 0.5 s, 1 s, 2 s, 4 s, and 8 s, respectively, throughout the trial duration. We defined an attentional modulation index (AMI) as,

$$\begin{aligned} AMI = corr(\hat{X}_A, X_{sp1}) - corr(\hat{X}_A, X_{sp2}) \\ + corr(\hat{X}_U, X_{sp2}) - corr(\hat{X}_U, X_{sp1}). \end{aligned} \tag{7.3}$$

A positive value of this index suggests that speaker 1 is the attended speaker, and a negative value votes speaker 2 to be the attended speaker for this window. Decoding accuracy is defined as the

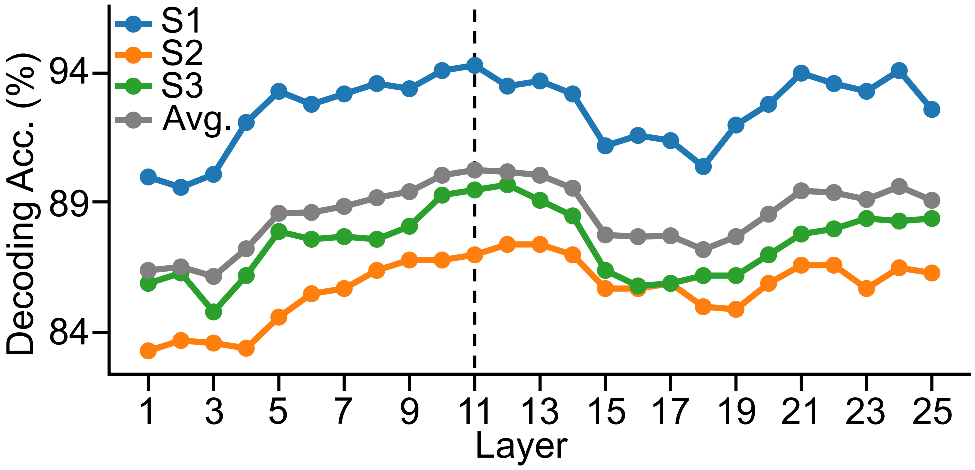


Figure 7.2: Accuracy of attention decoding using representations from each layer of WavLM with a 4-second decoding window. The 11-th layer shows the best average performance across subjects.

percentage of windows that were correctly classified.

7.3 Results and Discussion

7.3.1 Decoding Accuracy for Each Layer of the SSL Model

Fig. 7.2 shows the effect of different layer representations from WavLM on decoding accuracy. The accuracy improves as the layer depth increases, then slightly decreases before climbing again. The 11-th layer produces the best performance on average for the three subjects. The first layer, which is the output of the CNN encoder, extracts local features (~ 25 ms) from speech, resembling a spectrogram. The subsequent layers contain semantic information with more context. A recent layer-wise analysis of wav2vec 2.0 found an acoustic-linguistic hierarchy in layer-wise representation evolution, where shallow layers encode local acoustic information, followed by phonetics, word identity, and word meaning [204]. Therefore, Fig. 7.2 suggests that speech's higher-level features may be better decoded from the brain to enhance attention decoding accuracy. Pasad et al. [204] also noticed a reverse trend starting from the middle layer, which they attributed to the transformer layers' autoencoder-style behavior where deeper layers become closer to the input.

Table 7.1: Accuracy of attention decoding using various features extracted from clean speech (Avg. Over 3 Subjects, in %)

Feature	Decoding window size				
	0.5s	1s	2s	4s	8s
Envelope	63.3	71.6	79.5	86.0	91.3
Mel-spectrogram	65.6	72.3	80.8	88.5	91.5
WavLM	**72.9**	**78.7**	**85.2**	**90.3**	**92.6**
WavLM w/ causal ATT	72.2	78.5	84.6	89.4	92.3
WavLM w/ causal ATT & PE	72.0	77.9	84.1	89.1	92.5
WavLM w/ random init.	62.8	68.4	74.1	79.8	87.1

7.3.2 Decoding Accuracy for Different Reconstruction Targets

Table 7.1 compares the results of different features used for auditory attention decoding. The acoustic features envelope and 28-basis Mel-spectrogram are the baseline features. The envelope and Mel-spectrogram features were Z-scored before training and inference. Results show that all the features extracted from WavLM consistently outperform the baseline (paired *t*-test, $p < 0.001$ for win sizes 0.5 s, 1 s, 2 s, and 4 s; $p < 0.05$ for win size 8 s). WavLM performs especially well compared to the baseline when the decoding window size is small. The causal configuration resulted in a slight performance decrease, but it is expected that further fine-tuning can reduce this decrease. A control experiment using WavLM with random initialization shows significantly worse results than the baseline, confirming that the performance gain for WavLM was due to self-supervised learned representations, not due to its architecture or feature dimension.

Because clean speech of individual speakers is usually unavailable, we used an automatic speech separation model introduced in Section 2.3 to separate the mixed speech. Results in Table 7.2 show a slight decrease in accuracy compared to those in Table 1 due to imperfect speech separation, but this difference is small and all features are similarly affected. Despite this, WavLM remains superior to speech envelope and Mel-spectrogram.

Table 7.2: Accuracy of attention decoding using various features extracted from separated speech (Avg. Over 3 Subjects, in %)

Feature	Decoding window size				
	0.5s	1s	2s	4s	8s
Envelope	63.0	70.1	78.5	85.2	90.1
Mel-spectrogram	64.6	71.1	79.2	86.3	90.4
WavLM	**72.2**	**77.8**	**83.9**	**88.6**	**92.4**
WavLM w/ causal ATT	70.8	76.6	82.5	88.1	92.1
WavLM w/ causal ATT & PE	70.6	76.1	82.3	88.1	91.7
WavLM w/ random init.	62.8	68.4	74.1	79.8	84.6

7.3.3 Number of Principle Components

We reduced the dimension of WavLM features using Principal Component Analysis (PCA). Table 7.3 presents the decoding accuracy for WavLM features with varying numbers of PCA components. Although accuracy decreases slightly with fewer components, with the same number of components, WavLM surpasses Mel-spectrogram notably.

Table 7.3: Accuracy of attention decoding using WavLM feature with various component numbers.

Feature	Decoding window size				
	0.5s	1s	2s	4s	8s
Mel-spectrogram (28 dims)	64.6	71.1	79.2	86.3	90.4
WavLM causal (1024 dims)	70.6	76.1	82.3	88.1	91.7
200 PCs	70.3	75.7	82.2	88.1	91.7
100 PCs	70.0	75.5	81.9	87.8	91.7
50 PCs	69.5	75.2	81.8	87.4	91.4
28 PCs	69.0	74.8	81.5	87.7	91.7

7.3.4 Dynamic Switching of Attention

We simulated dynamic attention switching by concatenating the first 10 seconds and last 10 seconds of neural responses in each trial where the subject was attending to Spk1 in the first 10 s and switched to Spk2 afterward. We calculated AMI scores for WavLM w/ causal ATT & PE and

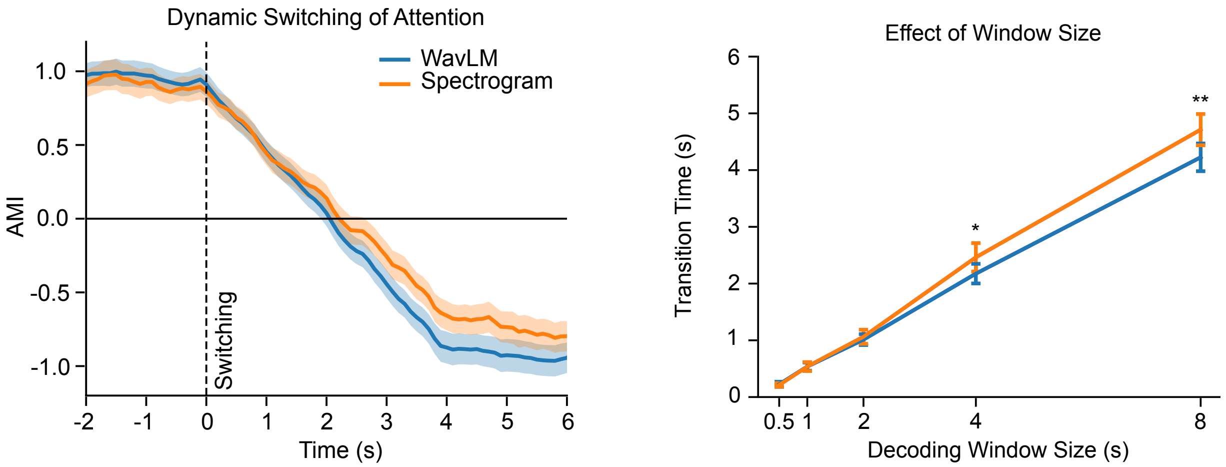


Figure 7.3: (1) The upper plot displays attention switching from speaker 1 to speaker 2. The dashed line represents the switch in attention. The average AMI for the three subjects is shown with a 4-second decoding window. (2) The bottom plot shows the transition time for detecting a switch which was measured as the moment when AMI crosses zero. Statistical significance is indicated by asterisks: $\star$ for $p < 0.05$ and $\star\star$ for $p < 0.01$.

Mel-spectrogram, respectively, using a sliding window of 4s. The upper panel of Fig. 7.3 shows the AMI scores averaged over all the subjects and trials. The averaged AMI scores were scaled between -1 and 1. WavLM and spectrogram exhibit a similar pattern but WavLM detects the switch faster. The bottom panel of Fig. 7.3 shows the average transition times for five different sliding window durations. As expected, the transition times increase for longer durations. There are no significant differences between WavLM and spectrogram for window size 2 s and below (paired t-test, $p > 0.2$); However, WavLM has a shorter transition time for window sizes 4 s ($p < 0.05$) and 8 s ($p < 0.01$).

7.3.5 Comparison of Features in Predicting Neural Activity

To gain further insight into why SSL features provide a higher neural decoding accuracy, we used a forward model to predict the response of single neural sites from different layers of WavLM. While the stimulus reconstruction method uses a backward model, here we trained forward models, spatiotemporal filters, G_R, that predict neural activity based on various stimulus features. The

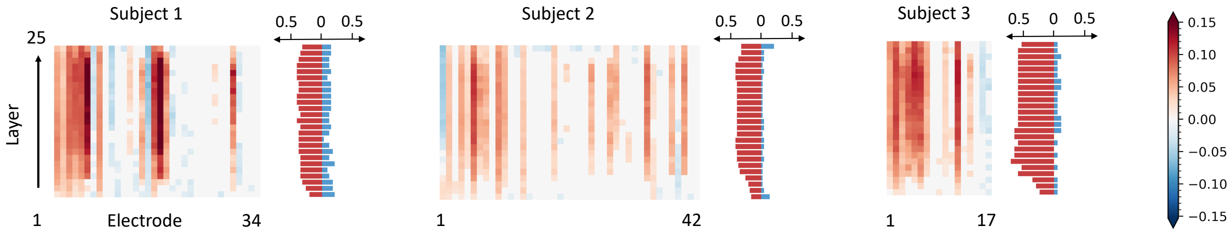


Figure 7.4: The improvement in r-value between the actual and predicted neural activity using WavLM features compared to using spectrogram for each layer and electrode. A positive value indicates WavLM features provide a better prediction of the neural activity at that electrode than the spectrogram, while a negative value indicates that the spectrogram is more accurate. Zeros (show in white) indicate no significant difference between the two features ($p > 0.05$). The bar plots show the proportion of electrodes that are more responsive to WavLM features (red) and to spectrogram (blue), respectively.

mathematical principle is similar to that of stimulus reconstruction, just in the opposite direction,

$$\hat{R}(e,t) = \sum_{e} \sum_{\tau} G_R(e,n,\tau) X_A(n,t-\tau), \tag{7.4}$$

where the time lag τ ranges from 0 to 200 ms. We measured the correlation (r-value) between the reconstructed and actual neural activity for each electrode. We assessed the improvement in r-value using each layer of WavLM compared to the Mel-spectrogram, where we utilized the first 100 PCs of WavLM features. If the results of a paired t-test showed no statistical difference between using WavLM features and the Mel-spectrogram (with a $p > 0.05$), the improvement value was set to zero. Additionally, we calculated the percentage of electrodes that were more accurately predicted using spectrogram and the percentage of electrodes that were more accurately predicted using WavLM features for each layer.

The results for the three subjects are presented in Fig. 7.4. The middle layers of WavLM generally provided better predictions compared to the shallow layers and the deepest layers. Although some electrodes were more accurately predicted using the acoustic spectrogram (shown in blue), a larger proportion of electrodes were better predicted using WavLM features (shown in red). The results in Fig. 7.4 indicate that different regions of the auditory cortex encode different levels of speech information, inspiring combining different layers of WavLM features to further improve

AAD accuracy.

7.4 Conclusion

This chapter investigates the use of self-supervised speech representations to enhance attentional decoding in multi-talker situations. Results show that substituting traditional speech features with latent features from WavLM result in improved attention decoding accuracy and speed, paving the path to more swift brain-controlled hearing devices. These findings suggest the need for further exploration of self-supervised speech representations in auditory neural decoding and their potential to improve our understanding of how the human brain makes attentional selections.

Conclusion and Future Work

In this dissertation, we designed speech separation models for brain-controlled hearing technologies. We used deep learning methods and investigated multiple prevalent challenges, including improving signal quality of separated speech, improving model robustness against adverse environments with background noise, room reverberation, and source motion, and improving the models' generalization capabilities for real-world recordings. In addition to these challenges, we considered key factors for hearing device development, such as designing separation models with causal configurations for low-latency separation and designing binaural separation models for preserving spatial the cues of individual speakers. A critical part of this dissertation is the integration of speech separation with auditory attention decoding, which we refer to as SS-AAD systems. We first designed a basic AAD experiment to validate the concept. Subsequently, we progressed to developing a realistic AAD paradigm that replicates the complex acoustics of real-world environments. This step is essential for bringing SS-AAD systems closer to real-world use. The evaluation results show that the proposed SS-AAD systems enhance subjective and objective quality of perceiving the attended speaker and reduce listening effort in a multi-talker mixture. By combining the advances in automatic speech separation and brain-computer interfaces, this dissertation provides a solution to brain-controlled hearing that can dramatically improve the quality of life for the hearing-impaired and augment hearing capabilities for the general public.

Below I outline a few open questions and future directions.

Better separation performance for human listening. Recent years have witnessed significant progress in deep neural network-based speech separation models. The scale-invariant signal-to-distortion ratio improvement (SI-SDRi) of the leading model has seen a remarkable rise, from 10.8 dB in 2015 [20] to 23.9 dB in 2023 [205], on the WSJ0-2mix dataset. Currently, most separation and enhancement models are trained using SI-SDR or similar metrics, such as SNR, SI-SNR and MSE, to optimize speech signal quality. However, an increased signal quality does not necessarily lead to improved human auditory perception. In Chapter 6, we revealed that while deep neural networks excelled in noise reduction due to the power of nonlinear models, they offered only marginal improvements in speech intelligibility and were rated poorly in human listening tests. To investigate the potential of deep neural networks for hearing devices, we must rethink both the training objectives and objective evaluation metrics. Luo et al. [206] designed auxiliary autoencoding training (A2T) to control the distortion on the direct-path signals and improve the recognition accuracy in reverberant separation. Adversarial loss [207, 208] has also been explored to approximate the distribution of clean speech. Moreover, recent studies have started optimizing speech separation and enhancement models using perceptual-related measures such as STOI [209, 210], PESQ [211, 212, 213], and human-assessed MOS [214]. However, there remains a lack of comprehensive studies exploring the extent to which perceptually-related objectives improve models for human listening. The exploration of other novel objectives to optimize speech separation models for better human listening experience continues to be an important direction of research.

Multi-modal speech separation. In this dissertation, our focus is on speech separation using audio signals alone. However, there have been extensive studies improving speech separation models by adding visual information, such as lip motion and face features [215, 216, 217]. These visual cues, correlated to speech content and speaker characteristics, usually remain consistent across different acoustic environments and, therefore, could benefit speech separation in challenging acoustic environments like cocktail parties. Another emerging multi-modal approach is text-informed speech separation [218, 219, 220]. In the context of hearable devices, this is termed as semantic hearing [221] where users can use semantic instructions to guide devices in focusing on or filtering out specific sounds in real-world environments. With the advent of large language models (LLMs) [222], text-promptable models have become a popular topic across various fields. There is a growing trend of using natural language to instruct models. It has been shown a simple textual description of the desired audio source can inform the separation model to isolate that source [223, 224]. We anticipate in the future users will effortlessly communicate with smart hearing aids using typed instructions or voice commands. Users can also directly modify acoustic scenes to their preference directly. Interesting questions arise regarding how to involve these advancements with brain-computer interface technologies. The integration of brain signals with these novel, multi-modal approaches will open up the possibility of developing smart hearing aids that are more intuitive and responsive to the user's auditory needs and preferences. Moreover, identifying the joint semantic space between brain activities and these new modalities will also be an area of interest.

On-device models. Deploying speech separation models to devices requires low latency, minimal computational complexity, lightweight model structures, and low power consumption. We have investigated efficient model architectures that decrease the model size and complexity without sacrificing the performance [225, 226]. Continued development of efficient and effective models for hearing aids remains a critical direction. A noteworthy trend in recent years is training large models on large datasets. The supervised model, Whisper, with 1.6 billion parameters, is trained on 680,000 hours of data [227]. Similarly, self-supervised speech representation learning models wav2vec [197], HuBERT [198], wavLM [199] with billions of parameters are trained on substantial volumes of unlabeled data and have been used as an upstream model in many speech tasks. Audio foundation model is an emerging topic, aiming to unify common audio and speech tasks in a generative framework. Drawing inspiration from the successes of large language models [228], there is an expectation of similar emergent capabilities from large-scale audio models. However, the direct application of these large models in hearing aids is impractical due to their size and complexity. Therefore, future research should look into ways of adapting the capabilities of these advanced models for use in on-device applications.

Closed-loop AAD. Most AAD studies, including the one in this dissertation, have used an open loop setup where the data was recorded and analyzed offline, and the audio that was delivered to the listener was not modulated by attention. A realistic brain-controlled hearing device requires real-time closed-loop implementation. This would require all the blocks of the proposed framework, including speech separation and AAD to work together synchronously and be implemented in a causal real-time manner. Moreover, how to best manipulate/re-mix the acoustic scene once the attended talker has been decoded also requires further investigation. A recent closed-loop AAD study [229] using scalp EEG found that it is imperative for the system to quickly track switches in attention in order to achieve desirable user experience. Fast detection of attention switches requires shorter window durations, made possible with invasive EEG techniques. A closed-loop version of our study, in which the attended talker is enhanced in an online real-time fashion and fed back to the subject, will be an important research direction in the future.

References

[1] M. Slaney *et al.*, "Auditory measures for the next billion users," *Ear and Hearing*, vol. 41, 131S–139S, 2020.

[2] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 342–356, 2020.

[3] M. B. Hoy, "Alexa, Siri, Cortana, and more: An introduction to voice assistants," *Medical Reference Services Quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[4] L. Barrault *et al.*, "SeamlessM4T-massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.

[5] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[6] Y. A. Li, C. Han, and N. Mesgarani, "StyleTTS: A style-based generative model for natural and diverse text-to-speech synthesis," *arXiv preprint arXiv:2205.15439*, 2022.

[7] Y. A. Li, C. Han, V. S. Raghavan, G. Mischler, and N. Mesgarani, "StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," in *Advances in Neural Information Processing Systems*, 2023.

[8] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[9] V Hamacher *et al.*, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 18, pp. 1–15, 2005.

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

[11] R. Plomp, "Noise, amplification, and compression: Considerations of three main issues in hearing aid design," *Ear and Hearing*, vol. 15, no. 1, pp. 2–12, 1994.

[12] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[13] N. Ding and J. Z. Simon, “Emergence of neural encoding of auditory objects while listening to competing speakers,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11 854–11 859, 2012.

[14] J. A. O’Sullivan *et al.*, “Attentional selection in a cocktail party environment can be decoded from single-trial eeg,” *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[15] I. Kavalerov *et al.*, “Universal sound separation,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2019, pp. 175–179.

[16] M. N. Schmidt and R. K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization.,” in *Proc. Interspeech*, Citeseer, vol. 2, 2006, pp. 2–5.

[17] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, “Discriminative NMF and its application to single-channel source separation.,” in *Proc. Interspeech*, 2014, pp. 865–869.

[18] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, “Probabilistic modeling paradigms for audio source separation,” in *Machine Audition: Principles, Algorithms and Systems*, IGI global, 2011, pp. 162–185.

[19] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[20] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2016, pp. 31–35.

[21] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, “Single-channel multi-speaker separation using deep clustering,” in *Proc. Interspeech*, 2016, pp. 545–549.

[22] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2017, pp. 241–245.

[23] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[24] J. Agnew and J. M. Thornton, “Just noticeable and objectionable group delays in digital hearing aids,” *Journal of the American Academy of Audiology*, vol. 11, no. 06, pp. 330–336, 2000.

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

[26] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1075–1084, 2017.

[27] Q. Liu, Y. Xu, P. J. Jackson, W. Wang, and P. Coleman, "Iterative deep neural networks for speaker-independent binaural blind speech separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2018, pp. 541–545.

[28] P. Dadvar and M. Geravanchizadeh, "Robust binaural speech separation in adverse conditions based on deep neural network with modified spatial features and training target," *Speech Communication*, vol. 108, pp. 41–52, 2019.

[29] M. Sams, M. Hämäläinen, R. Hari, and L. McEvoy, "Human auditory cortical mechanisms of sound lateralization: I. Interaural time differences within sound," *Hearing Research*, vol. 67, no. 1-2, pp. 89–97, 1993.

[30] T. C. Yin, "Neural mechanisms of encoding binaural localization cues in the auditory brainstem," in *Integrative Functions in the Mammalian Auditory Pathway*, Springer, 2002, pp. 99–159.

[31] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, p. 063 297, 2006.

[32] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Robustness analysis of binaural hearing aid beamformer algorithms by means of objective perceptual quality measures," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2007, pp. 315–318.

[33] K. Reindl, Y. Zheng, and W. Kellermann, "Analysis of two generic Wiener filtering concepts for binaural speech enhancement in hearing aids," in *Proc. European Signal Processing Conference (EUSIPCO)*, IEEE, 2010, pp. 989–993.

[34] J. I. Marin-Hurtado, D. N. Parikh, and D. V. Anderson, "Perceptually inspired noise-reduction method for binaural hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1372–1382, 2011.

[35] M. Azarpour and G. Enzner, "Binaural noise reduction via cue-preserving MMSE filter and adaptive-blocking-based noise PSD estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, no. 1, pp. 1–17, 2017.

[36] M. Zohourian and R. Martin, "GSC-based binaural speaker separation preserving spatial cues," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2018, pp. 516–520.

[37] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[38] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2449–2464, 2015.

[39] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural cue preservation using multi-channel wiener filtering and interaural transfer functions," in *Proc. International Workshop Acoustic Echo Noise Control (IWAENC)*, 2006, pp. 1–4.

[40] C. Han, Y. Luo, and N. Mesgarani, "Real-time binaural speech separation with preserved spatial cues," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6404–6408.

[41] C. Han, Y. Luo, and N. Mesgarani, "Binaural speech separation of moving speakers with preserved spatial cues.," in *Proc. Interspeech*, 2021, pp. 3505–3509.

[42] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2020, pp. 46–50.

[43] Y. Luo, Z. Chen, C. Han, C. Li, T. Zhou, and N. Mesgarani, "Rethinking the separation layers in speech separation networks," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2021, pp. 1–5.

[44] C. Li *et al.*, "Dual-path modeling for long recording speech separation in meetings," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2021, pp. 5739–5743.

[45] C. Li *et al.*, "Dual-path RNN for long recording speech separation," in *Proc. Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 865–872.

[46] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2021, pp. 21–25.

[47] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[48] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2014, pp. 4052–4056.

[49] K. Žmolíková *et al.*, “SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[50] M. Delcroix *et al.*, “Improving speaker discrimination of target speech extraction with time-domain speakerbeam,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2020, pp. 691–695.

[51] K. Veselỳ, S. Watanabe, K. Žmolíková, M. Karafiát, L. Burget, and J. H. Černockỳ, “Sequence summarizing neural network for speaker adaptation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2016, pp. 5315–5319.

[52] Q. Wang *et al.*, “VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking,” in *Proc. Interspeech*, 2019, pp. 2728–2732.

[53] X. Xiao *et al.*, “Single-channel speech extraction using speaker inventory and attention network,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2019, pp. 86–90.

[54] C. Han *et al.*, “Continuous speech separation using speaker inventory for long recording,” *Proc. Interspeech*, p. 5, 2021.

[55] C. Han and N. Mesgarani, “Online binaural speech separation of moving speakers with a Wavesplit network,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[56] P. Seetharaman, G. Wichern, J. Le Roux, and B. Pardo, “Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2019, pp. 356–360.

[57] L. Drude, D. Hasenklever, and R. Haeb-Umbach, “Unsupervised training of a deep clustering model for multichannel blind source separation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2019, pp. 695–699.

[58] M. Togami, Y. Masuyama, T. Komatsu, and Y. Nakagome, “Unsupervised training for deep speech source separation with Kullback-Leibler divergence based probabilistic loss function,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2020, pp. 56–60.

[59] Y. Nakagome, M. Togami, T. Ogawa, and T. Kobayashi, “Mentoring-reverse mentoring for unsupervised multi-channel speech source separation.,” in *Proc. Interspeech*, 2020, pp. 86–90.

[60] C. Han, E. M. Kaya, K. Hoefer, M. Slaney, and S. Carlile, “Multi-channel speech denoising for machine ears,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022, pp. 276–280.

[61] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, “Unsupervised sound separation using mixture invariant training,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 3846–3857.

[62] A. Sivaraman, S. Wisdom, H. Erdogan, and J. R. Hershey, “Adapting speech separation to real-world meetings using mixture invariant training,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022, pp. 686–690.

[63] C. Han, K. Wilson, S. Wisdom, and J. R. Hershey, “Unsupervised multi-channel separation and adaptation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2024.

[64] M. Hosseini, L. Celotti, and E. Plourde, “End-to-end brain-driven speech enhancement in multi-talker conditions,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1718–1733, 2022.

[65] J. Zhang, Q. Xu, Q.-S. Zhu, and Z.-H. Ling, “BASEN: Time-domain brain-assisted speech enhancement network with convolutional cross attention in multi-talker conditions,” in *Proc. Interspeech*, 2023, pp. 3117–3121.

[66] Z. Pan, M. Borsdorf, S. Cai, T. Schultz, and H. Li, “NeuroHeed: Neuro-steered speaker extraction using eeg signals,” *arXiv preprint arXiv:2307.14303*, 2023.

[67] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor, “Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech,” *Current Biology*, vol. 28, no. 5, pp. 803–809, 2018.

[68] N. Das, T. Francart, and A. Bertrand, “Auditory attention detection dataset KULeuven,” *Zenodo*, 2019.

[69] S. A. Fuglsang, J. Märcher-Rørsted, T. Dau, and J. Hjortkjær, "Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention," *Journal of Neuroscience*, vol. 40, no. 12, pp. 2562–2572, 2020.

[70] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: Implications for online, daily-life applications," *Journal of Neural Engineering*, vol. 12, no. 4, p. 046 007, 2015.

[71] L. Fiedler, M. Wöstmann, C. Graversen, A. Brandmeyer, T. Lunner, and J. Obleser, "Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech," *Journal of Neural Engineering*, vol. 14, no. 3, p. 036 020, 2017.

[72] S. L. Metzger *et al.*, "A high-performance neuroprosthesis for speech decoding and avatar control," *Nature*, vol. 620, no. 7976, pp. 1037–1046, 2023.

[73] F. R. Willett *et al.*, "A high-performance speech neuroprosthesis," *Nature*, vol. 620, no. 7976, pp. 1031–1036, 2023.

[74] E. Ceolini *et al.*, "Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception," *NeuroImage*, vol. 223, p. 117 282, 2020.

[75] Z. Pan, G. Wichern, F. G. Germain, S. Khurana, and J. L. Roux, "NeuroHeed+: Improving neuro-steered speaker extraction with joint auditory attention detection," *arXiv preprint arXiv:2312.07513*, 2023.

[76] C. Han, Y. Luo, and N. Mesgarani, "Online deep attractor network for real-time single-channel speech separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2019, pp. 361–365.

[77] C. Han, J. O'Sullivan, Y. Luo, J. Herrero, A. D. Mehta, and N. Mesgarani, "Speaker-independent auditory attention decoding without access to clean speech sources," *Science Advances*, vol. 5, no. 5, eaav6134, 2019.

[78] C. Han, V. Choudhari, Y. A. Li, and N. Mesgarani, "Improved decoding of attentional selection in multi-talker environments with self-supervised learned speech representation," in *Proc. in IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2023.

[79] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2017, pp. 246–250.

[80] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[81] L. Bottou and Y. Bengio, “Convergence properties of the k-means algorithms,” in *Advances in Neural Information Processing Systems*, 1995, pp. 585–592.

[82] E. Liberty, R. Sriharsha, and M. Sviridenko, “An algorithm for online k-means clustering,” in *Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, SIAM, 2016, pp. 81–89.

[83] G. Strang, “Introduction to linear algebra, 3rd edition,” *Wellesley-Cambridge Press*, 1993.

[84] T. K. Moon, “The expectation-maximization algorithm,” *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[85] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[86] D. B. Paul and J. Baker, “The design for the Wall Street Journal-based CSR corpus,” in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[87] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, vol. 2, 2001, pp. 749–752.

[88] J. Jensen and C. H. Taal, “An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[89] A. Alinaghi, W. Wang, and P. J. Jackson, “Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2013, pp. 684–688.

[90] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, “The CIPIC HRTF database,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2001, pp. 99–102.

[91] X. Sun, R. Xia, J. Li, and Y. Yan, “A deep learning based binaural speech enhancement approach with spatial cues preservation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2019, pp. 5766–5770.

[92] H. Haas, “The influence of a single echo on the audibility of speech,” *The Journal of the Audio Engineering Society*, vol. 20, no. 2, pp. 146–159, 1972.

[93] D. Stoller, S. Ewert, and S. Dixon, “Wave-U-Net: A multi-scale neural network for end-to-end audio source separation,” in *Proc. International Society for Music Information Retrieval ISMIR*, 2018, pp. 334–340.

[94] S. Venkataramani, J. Casebeer, and P. Smaragdis, “End-to-end source separation with adaptive front-ends,” in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, IEEE, 2018, pp. 684–688.

[95] L. Zhang, Z. Shi, J. Han, A. Shi, and D. Ma, “FurcaNeXt: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks,” in *International Conference on Multimedia Modeling*, Springer, 2020, pp. 653–665.

[96] R. Gu *et al.*, “End-to-end multi-channel speech separation,” *arXiv preprint arXiv:1905.06286*, 2019.

[97] Y. Luo, E. Ceolini, C. Han, S.-C. Liu, and N. Mesgarani, “FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2019.

[98] J. Thiemann, N. Ito, and E. Vincent, “The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings,” in *Proceedings of Meetings on Acoustics*, vol. 19, 2013.

[99] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[100] A Koutvas, E. Dermatas, and G. Kokkinakis, “Blind speech separation of moving speakers in real reverberant environments,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, vol. 2, 2000, pp. II1133–II1136.

[101] R. Mukai, H. Sawada, S. Araki, and S. Makino, “Real-time blind source separation for moving speech signals,” in *Speech Enhancement*, Springer, 2005, pp. 353–369.

[102] J. Zhang and P.-C. Ching, “Blind separation of moving speech sources using short-time LOD based ICA method,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, vol. 3, 2007, pp. III–957.

[103] R. Mukai, H. Sawada, S. Araki, and S. Makino, “Robust real-time blind source separation for moving speakers in a room,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, vol. 5, 2003, pp. V–469.

[104] R. Mukai, H. Sawada, S. Araki, and S. Makino, “Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction,” *IEICE Transac-*

tions on Fundamentals of Electronics, Communications and Computer Sciences, vol. 87, no. 8, pp. 1941–1948, 2004.

[105] N. Chong, S. Nordholm, B. T. Vo, and I. Murray, "Tracking and separation of multiple moving speech sources via cardinality balanced multi-target multi bernoulli (CBMeMBer) filter and time frequency masking," in *Proc. IEEE International Conference on Control, Automation and Information Sciences (ICCAIS)*, IEEE, 2016, pp. 88–93.

[106] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 281–295, 2017.

[107] T. Higuchi, N. Takamune, T. Nakamura, and H. Kameoka, "Underdetermined blind separation and tracking of moving sources based ONDOA-HMM," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2014, pp. 3191–3195.

[108] M. Taseska and E. A. Habets, "Blind source separation of moving sources using sparsity-based source detection and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 657–670, 2017.

[109] X. Li *et al.*, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1997–2012, 2017.

[110] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[111] K. Tan, B. Xu, A. Kumar, E. Nachmani, and Y. Adi, "SAGRNN: Self-attentive gated RNN for binaural speaker separation with interaural cue preservation," *IEEE Signal Processing Letters*, 2020.

[112] Z. Chen *et al.*, "Continuous speech separation: Dataset and analysis," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7284–7288.

[113] W. Zhang *et al.*, "Separating long-form speech with group-wise permutation invariant training," in *Proc. Interspeech*, 2022, pp. 5383–5387.

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

[115] T. Zhou, Y. Zhao, J. Li, Y. Gong, and J. Wu, "CNN with phonetic attention for text-independent speaker verification," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2019, pp. 718–725.

[116] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *AAAI*, 2018.

[117] J. Wang *et al.*, "Deep extractor network for target speaker recovery from single channel speech mixtures," in *Proc. Interspeech*, 2018, pp. 307–311.

[118] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "A unified framework for neural speech separation and extraction," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6975–6979.

[119] P. Wang *et al.*, "Speech separation using speaker inventory," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2019, pp. 230–236.

[120] A. Janin *et al.*, "The ICSI meeting corpus," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, vol. 1, 2003, pp. I–I.

[121] J. Carletta *et al.*, "The AMI meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*, Springer, 2005, pp. 28–39.

[122] T. Yoshioka *et al.*, "Advances in online audio-visual meeting transcription," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2019, pp. 276–283.

[123] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. Interspeech*, 2018, pp. 1561–1565.

[124] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5739–5743.

[125] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[126] Ö. Çetin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition," in *9th International Conference on Spoken Language Processing*, 2006.

[127] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[128] A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: A large-scale speaker identification dataset,” in *Proc. Interspeech*, 2017, pp. 2616–2620.

[129] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

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

[131] I. Cohen and B. Berdugo, “Speech enhancement for non-stationary noise environments,” *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[132] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2013.

[133] F. Weninger *et al.*, “Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR,” in *International Conference on Latent Variable Analysis and Signal Separation*, Springer, 2015, pp. 91–99.

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

[135] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, “Unsupervised sound separation using mixtures of mixtures,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 3846–3857.

[136] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, “Improved MVDR beamforming using single-channel mask prediction networks,” in *Proc. Interspeech*, 2016, pp. 1981–1985.

[137] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, “Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2017, pp. 286–290.

[138] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, “Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6384–6388.

[139] Z.-Q. Wang, P. Wang, and D. Wang, “Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2001–2014, 2021.

[140] Z.-Q. Wang *et al.*, "Sequential multi-frame neural beamforming for speech separation and enhancement," in *Proc. Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 905–911.

[141] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. European Signal Processing Conference (EUSIPCO)*, IEEE, 2016, pp. 1153–1157.

[142] J. T. Kent, "Data analysis for shapes and images," *The Journal of Statistical Planning and Inference*, vol. 57, no. 2, pp. 181–193, 1997.

[143] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, 2007, pp. 3247–3250.

[144] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[145] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2009.

[146] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[147] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on Amazon Mechanical Turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, 2010.

[148] B Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[149] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[150] E. Tzinis, S. Venkataramani, and P. Smaragdis, "Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2019, pp. 81–85.

[151] L. Drude, D. Hasenklever, and R. Haeb-Umbach, "Unsupervised training of a deep clustering model for multichannel blind source separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 695–699.

[152] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, “End-to-end microphone permutation and number invariant multi-channel speech separation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 6394–6398.

[153] T. Yoshioka *et al.*, “VarArray: Array-geometry-agnostic continuous speech separation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022, pp. 6027–6031.

[154] S. Wisdom *et al.*, “Differentiable consistency constraints for improved deep speech enhancement,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2019, pp. 900–904.

[155] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR–half-baked or well done?” In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2019, pp. 626–630.

[156] R. Carhart and T. W. Tillman, “Interaction of competing speech signals with hearing losses,” *Archives of Otolaryngology*, vol. 91, no. 3, pp. 273–279, 1970.

[157] B. D. Van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[158] S. Van Eyndhoven, T. Francart, and A. Bertrand, “EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 5, pp. 1045–1056, 2016.

[159] A. Aroudi, D. Marquardt, and S. Daclo, “EEG-based auditory attention decoding using steerable binaural superdirective beamformer,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2018, pp. 851–855.

[160] J. O’Sullivan *et al.*, “Neural decoding of attentional selection in multi-speaker environments without access to clean sources,” *Journal of Neural Engineering*, vol. 14, no. 5, p. 056 001, 2017.

[161] C. Destrieux, B. Fischl, A. Dale, and E. Halgren, “Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature,” *NeuroImage*, vol. 53, no. 1, pp. 1–15, 2010.

[162] E. M. Z. Golumbic *et al.*, “Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”,” *Neuron*, vol. 77, no. 5, pp. 980–991, 2013.

[163] N. E. Crone, D. Boatman, B. Gordon, and L. Hao, “Induced electrocorticographic gamma activity during auditory perception,” *Clinical Neurophysiology*, vol. 112, no. 4, pp. 565–582, 2001.

[164] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, “Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex,” *Journal of Neurophysiology*, vol. 102, no. 6, pp. 3329–3339, 2009.

[165] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, “Towards reconstructing intelligible speech from the human auditory cortex,” *Scientific Reports*, vol. 9, no. 1, p. 874, 2019.

[166] D. S. Brungart, “Informational and energetic masking effects in the perception of two simultaneous talkers,” *The Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1101–1109, 2001.

[167] ITU-T Recommendation, “Vocabulary for performance and quality of service,” *The International Telecommunication Union Telecommunication Standardization Sector (ITU-T)*, 2006.

[168] S. Ray and J. H. Maunsell, “Different origins of gamma rhythm and high-gamma activity in macaque visual cortex,” *PLOS Biology*, vol. 9, no. 4, e1000610, 2011.

[169] P. W. Hullett, L. S. Hamilton, N. Mesgarani, C. E. Schreiner, and E. F. Chang, “Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli,” *Journal of Neuroscience*, vol. 36, no. 6, pp. 2014–2026, 2016.

[170] K. Ovtcharov, O. Ruwase, J.-Y. Kim, J. Fowers, K. Strauss, and E. Chung, *Accelerating deep convolutional neural networks using specialized hardware*, 2015.

[171] R. Andri, L. Cavigelli, D. Rossi, and L. Benini, “YodaNN: An ultra-low power convolutional neural network accelerator based on binary weights,” in *Proc. IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, IEEE, 2016, pp. 236–241.

[172] G. Lacey, G. W. Taylor, and S. Areibi, “Deep learning on FPGAs: Past, present, and future,” *arXiv preprint arXiv:1602.04283*, 2016.

[173] S. A. Fuglsang, T. Dau, and J. Hjortkjær, “Noise-robust cortical tracking of attended speech in real-world acoustic scenes,” *NeuroImage*, vol. 156, pp. 435–444, 2017.

[174] A. de Cheveigné, D. D. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, and E. Lalor, “Decoding the auditory brain with canonical component analysis,” *NeuroImage*, vol. 172, pp. 206–216, 2018.

[175] S. Akram, A. Presacco, J. Z. Simon, S. A. Shamma, and B. Babadi, “Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling,” *NeuroImage*, vol. 124, pp. 906–917, 2016.

[176] S. Miran, S. Akram, A. Sheikhattar, J. Z. Simon, T. Zhang, and B. Babadi, “Real-time tracking of selective auditory attention from M/EEG: A bayesian filtering approach,” *Frontiers in Neuroscience*, vol. 12, p. 262, 2018.

[177] P. Patel, L. K. Long, J. L. Herrero, A. D. Mehta, and N. Mesgarani, “Joint representation of spatial and phonetic features in the human core auditory cortex,” *Cell Reports*, vol. 24, no. 8, pp. 2051–2062, 2018.

[178] S. Geirnaert *et al.*, “Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices,” *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, 2021.

[179] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, “EEG-based detection of the locus of auditory attention with convolutional neural networks,” *Elife*, vol. 10, e56481, 2021.

[180] B. J. Borgström, M. S. Brandstein, G. A. Ciccarelli, T. F. Quatieri, and C. J. Smalt, “Speaker separation in realistic noise environments with applications to a cognitively-controlled hearing aid,” *Neural Networks*, vol. 140, pp. 136–147, 2021.

[181] G. Ciccarelli *et al.*, “Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods,” *Scientific Reports*, vol. 9, no. 1, p. 11 538, 2019.

[182] G. Kidd, T. L. Arbogast, C. R. Mason, and F. J. Gallun, “The advantage of knowing where to listen,” *The Journal of the Acoustical Society of America*, vol. 118, no. 6, pp. 3804–3815, 2005.

[183] P. Patel, K. van der Heijden, S. Bickel, J. L. Herrero, A. D. Mehta, and N. Mesgarani, “Interaction of bottom-up and top-down neural mechanisms in spatial multi-talker speech perception,” *Current Biology*, vol. 32, no. 18, pp. 3971–3986, 2022.

[184] S. Makov, D. Pinto, P. H.-s. Yahav, L. M. Miller, and E. Z. Golumbic, ““unattended, distracting or irrelevant”: Theoretical implications of terminological choices in auditory selective attention research,” *Cognition*, vol. 231, p. 105 313, 2023.

[185] G. Marinato and D. Baldauf, “Object-based attention in complex, naturalistic auditory streams,” *Scientific Reports*, vol. 9, no. 1, p. 2854, 2019.

[186] W. K. Kirchner, “Age differences in short-term retention of rapidly changing information,” *Journal of Experimental Psychology*, vol. 55, no. 4, p. 352, 1958.

[187] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).*

[188] A. J. Spahr *et al.*, “Development and validation of the AzBio sentence lists,” *Ear and Hearing*, vol. 33, no. 1, p. 112, 2012.

[189] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using Kaldi,” in *Proc. Interspeech*, vol. 2017, 2017, pp. 498–502.

[190] M. Gorzel *et al.*, “Efficient encoding and decoding of binaural sound with resonance audio,” in *AES International Conference on Immersive and Interactive Audio*, Audio Engineering Society, 2019.

[191] J. A. Swets, *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Psychology Press, 2014.

[192] L. Kong *et al.*, “Auditory spatial attention representations in the human cerebral cortex,” *Cerebral Cortex*, vol. 24, no. 3, pp. 773–784, 2014.

[193] Y. Deng, I. Choi, and B. Shinn-Cunningham, “Topographic specificity of alpha power during auditory spatial attention,” *NeuroImage*, vol. 207, p. 116 360, 2020.

[194] K. Krumbholz *et al.*, “Representation of interaural temporal information from left and right auditory space in the human planum temporale and inferior parietal lobe,” *Cerebral Cortex*, vol. 15, no. 3, pp. 317–324, 2005.

[195] S. Haro, H. M. Rao, T. F. Quatieri, and C. J. Smalt, “Eeg alpha and pupil diameter reflect endogenous auditory attention switching and listening effort,” *European Journal of Neuroscience*, vol. 55, no. 5, pp. 1262–1277, 2022.

[196] A. Bednar and E. C. Lalor, “Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG,” *NeuroImage*, vol. 205, p. 116 283, 2020.

[197] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[198] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[199] S. Chen *et al.*, “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[200] S.-w. Yang *et al.*, "SUPERB: Speech processing universal performance benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[201] J. Millet *et al.*, "Toward a realistic model of speech processing in the brain with self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 428–33 443, 2022.

[202] A. R. Vaidya, S. Jain, and A. Huth, "Self-supervised models of audio effectively explain human cortical responses to speech," in *Proc. International Conference on Machine Learning (ICML)*, PMLR, 2022, pp. 21 927–21 944.

[203] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J.-R. King, "Decoding speech perception from non-invasive brain recordings," *Nature Machine Intelligence*, vol. 5, 1097–1107, 2023.

[204] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 914–921.

[205] S. Lutati, E. Nachmani, and L. Wolf, "Separate and diffuse: Using a pretrained diffusion model for improving source separation," *arXiv preprint arXiv:2301.10752*, 2023.

[206] Y. Luo, C. Han, and N. Mesgarani, "Distortion-controlled training for end-to-end reverberant speech separation with auxiliary autoencoding loss," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 825–832.

[207] Y. C. Subakan and P. Smaragdis, "Generative adversarial source separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2018, pp. 26–30.

[208] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. International Conference on Machine Learning (ICML)*, PMLR, 2019, pp. 2031–2041.

[209] H. Zhang, X. Zhang, and G. Gao, "Training supervised speech separation system to improve STOI and PESQ directly," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5374–5378.

[210] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-based source enhancement to increase objective sound quality assessment score," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1780–1792, 2018.

[211] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[212] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *Proceedings of the 10th international conference on Latent Variable Analysis and Signal Separation*, Springer, 2012, pp. 430–437.

[213] S.-W. Fu, C.-F. Liao, and Y. Tsao, "Learning with learned loss function: Speech enhancement with Quality-Net to improve perceptual evaluation of speech quality," *IEEE Signal Processing Letters*, vol. 27, pp. 26–30, 2019.

[214] K. M. Nayem and D. S. Williamson, "Attention-based speech enhancement using human quality perception modelling," *arXiv preprint arXiv:2303.13685*, 2023.

[215] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, 2018.

[216] J. Wu *et al.*, "Time domain audio visual speech separation," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2019, pp. 667–673.

[217] D. Michelsanti *et al.*, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

[218] L. Le Magoarou, A. Ozerov, and N. Q. Duong, "Text-informed audio source separation. Example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, vol. 79, no. 2, pp. 117–131, 2015.

[219] K. Schulze-Forster, C. S. Doire, G. Richard, and R. Badeau, "Joint phoneme alignment and text-informed speech separation on highly corrupted speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7274–7278.

[220] K. Kilgour, B. Gfeller, Q. Huang, A. Jansen, S. Wisdom, and M. Tagliasacchi, "Text-driven separation of arbitrary sounds," in *Proc. Interspeech*, 2022, pp. 5403–5407.

[221] B. Veluri, M. Itani, J. Chan, T. Yoshioka, and S. Gollakota, "Semantic hearing: Programming acoustic scenes with binaural hearables," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, ACM, 2023.

[222] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[223] X. Liu *et al.*, "Separate anything you describe," *arXiv preprint arXiv:2308.05037*, 2023.

[224] X. Hao, J. Wu, J. Yu, C. Xu, and K. C. Tan, "Typing to listen at the cocktail party: Text-guided target speaker extraction," *arXiv preprint arXiv:2310.07284*, 2023.

[225] Y. Luo, C. Han, and N. Mesgarani, “Ultra-lightweight speech separation via group communication,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2021, pp. 16–20.

[226] Y. Luo, C. Han, and N. Mesgarani, “Group communication with context codec for lightweight source separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1752–1761, 2021.

[227] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proc. International Conference on Machine Learning (ICML)*, PMLR, 2023, pp. 28 492–28 518.

[228] J. Wei *et al.*, “Emergent abilities of large language models,” *Transactions on Machine Learning Research*, 2022.

[229] A. Aroudi, E. Fischer, M. Serman, H. Puder, and S. Doclo, “Closed-loop cognitive-driven gain control of competing sounds using auditory attention decoding,” *Algorithms*, vol. 14, no. 10, p. 287, 2021.