DEEP NEURAL NETWORK-BASED APPROACHES FOR SINGLE-CHANNEL SPEAKER-CONDITIONED TARGET SPEAKER EXTRACTION

Von der Fakultät für Medizin und Gesundheitswissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels einer
Doktorin der Ingenieurwissenschaften (Dr.-Ing.)
angenommene Dissertation

von
Ragini Sinha
geboren am 15. Januar 1991
in Bihar (Indien)

Ragini Sinha: *Deep Neural Network-based Approaches for Single-channel Speaker-conditioned Target Speaker Extraction*

ERSTGUTACHTER:
Prof. Dr. ir. Simon Doclo, *Carl von Ossietzky Universität Oldenburg, Germany*

WEITERE GUTACHTER:
Prof. Dr. Bernd T. Meyer, *Carl von Ossietzky Universität Oldenburg, Germany*
Prof. Dr. ir. Emanuël Habets, *International Audio Laboratories Erlangen, Friedrich-Alexander Universität Erlangen-Nürnberg, Germany*

TAG DER DISPUTATION:
23. September 2025

ACKNOWLEDGMENTS

This thesis has been written at the Signal Processing Group in the Department of Medical Physics and Acoustics at the Carl von Ossietzky Universität Oldenburg in Oldenburg, Germany. I would like to take this opportunity to express my gratitude to everyone who supported and contributed to the successful completion of this thesis.

First and foremost, I would like to express my sincere gratitude to Simon Doclo for his constant support, insightful guidance, and invaluable advice throughout these years. His scientific insights and encouragement have been truly invaluable. I am also deeply thankful to Christian Rollwage for the many interesting discussions, his thoughtful and inspiring advice, and for the trust he placed in me, allowing me the freedom to explore my own ideas and research interests.

Furthermore, I would like to extend my sincere thanks to Emanuël Habets and Bernd T. Meyer for kindly reviewing this thesis and for their genuine interest in my work, which I greatly appreciate.

I would also like to thank Jan Rennies-Hochmuth for his continuous guidance and support throughout these years. My special thanks go to all members of the Signal Processing Group and Fraunhofer IDMT, Oldenburg, for creating such a collaborative and enjoyable working environment. In particular, I am sincerely grateful to Menno Müller, Mattes Ohlenbusch, and Paul Maria Reuter for their help and, even more, for all the fun, easygoing conversations and engaging discussions. I would also like to extend my thanks to Marvin Tammen, especially for his invaluable support during my early PhD days and for the many interesting discussions on recent research developments.

Finally, I would like to express my deepest gratitude to my family and friends for their continuous support and encouragement, especially to my parents, my younger sister, and my partner.

Saarbrücken, November 2025
Ragini Sinha

ABSTRACT

In everyday communication scenarios, such as meetings and social gatherings, undesired interfering speakers and background noise often degrade the quality and intelligibility of the desired target speaker. Various approaches have been developed to address this issue, such as blind source separation and speaker-conditioned target speaker extraction (SC-TSE). SC-TSE algorithms aim at extracting the desired speaker from the mixture by utilizing auxiliary information about the target speaker, such as reference speech, visual information, directional information, or speaker activity. A typical SC-TSE system consists of a speaker embedder network and a speaker separator network. The speaker embedder network generates target speaker-specific discriminative features from the auxiliary information, which guides the speaker separator network to extract the target speaker from the mixture. The aim of this thesis is to develop and evaluate novel DNN-based architectures to enhance the reliability, efficiency and robustness of single-channel SC-TSE algorithms utilizing reference speech as auxiliary information.

First, we propose three novel variants of long short-term memory (LSTM) cells for target speaker extraction in the time-frequency domain. These customized LSTM cells are specifically designed for the SC-TSE task, by optimizing how target speaker information is retained and updated within the LSTM cells. The first proposed variant customizes only the forget gate, enabling the selective retention of target speaker information while disregarding information from other sources in the mixture. The second proposed variant extends the first variant by customizing both the input and forget gates, enhancing the update mechanism of the cell state to reinforce target speaker-specific feature retention. The third proposed variant introduces an additional auxiliary-modulation gate within the LSTM cell, designed to dynamically learn both long-term and short-term speaker-specific feature discrimination. Experimental results on various mixture types show that all proposed variants of LSTM cells outperform standard LSTM cells in both unidirectional and bidirectional modes. The best performance is obtained using the auxiliary-gated LSTM cells, which yield scale-invariant signal-to-distortion ratio (SI-SDR) improvements up to 1.14 dB (unidirectional mode) and 1.09 dB (bidirectional mode) compared to standard LSTM cells.

Second, we propose two conformer-based architectures for target speaker extraction in the time domain. The first proposed architecture, Conformer-FFN, uses stacks of conformer and external feed-forward blocks, aiming at exploiting both local and global context features using conformer blocks, while reducing the overall number of parameters using external feed-forward blocks. The second proposed architecture, TCN-Conformer, uses stacks of temporal convolutional network (TCN) and conformer blocks, aiming at utilizing the best local context features using TCN

blocks and then exploiting both local and global context features using conformer blocks. Experimental results on various mixture types show that the proposed TCN-Conformer system outperforms the TCN-based baseline system and the proposed Conformer-FFN system. The best performance is obtained with four stacks of the TCN and conformer blocks, which yields SI-SDR improvements up to 2.64 dB over the TCN-based baseline and up to 3.44 dB over the Conformer-FFN system. To make the proposed TCN-Conformer system more suitable for real-time target speaker extraction, we replace the traditional multi-head self-attention (MHSA) in each conformer block of the speaker separator network with linear MHSA. Experimental results show that the TCN-Conformer system using linear MHSA outperforms the TCN-Conformer system using traditional MHSA, while achieving a significant reduction in computational cost and real-time factor. In addition, we show that using multi-condition training, it is possible to increase the robustness against background noise, reverberation and intrinsic variability (emotions) in the reference speech of the target speaker.

Third, we subjectively evaluate the performance of two SC-TSE algorithms by performing listening tests with normal-hearing (NH) and hearing-impaired (HI) listeners: an algorithm performing target speaker extraction using a real-valued mask in the time-frequency domain (Algo-1) and an algorithm performing target speaker extraction in the time domain using the proposed TCN-Conformer architecture (Algo-2). These algorithms were evaluated for challenging acoustic scenarios with up to two interfering speakers using three subjective evaluation methods: paired comparison, speech recognition thresholds (SRTs), and categorically scaled perceived listening effort. The results with fifteen NH and fifteen HI listeners show that Algo-2 significantly reduces listening effort, improves speech intelligibility, and is preferred compared to the unprocessed mixtures and Algo-1. Moreover, HI listeners experience greater benefits compared to NH listeners, e.g., in terms of listening effort, a reduction of 7-8 units (ESCU) compared to 4-5 units for NH listeners. For HI listeners with symmetric mild-to-moderate hearing loss, the results also suggest that hearing loss compensation is not necessary to obtain an algorithm benefit.

ZUSAMMENFASSUNG

In alltäglichen Kommunikationsszenarien, wie Besprechungen und gesellschaftlichen Zusammenkünften, beeinträchtigen unerwünschte Störsprecher und Hintergrundgeräusche häufig die Qualität und Verständlichkeit des gewünschten Zielsprechers. Verschiedene Ansätze wurden entwickelt, um dieses Problem zu lösen, darunter blinde Quellentrennung und speaker-conditioned target speaker extraction (SC-TSE). SC-TSE-Algorithmen zielen darauf ab, den gewünschten Sprecher aus dem Mischsignal zu extrahieren, indem sie zusätzliche Informationen über den Zielsprecher nutzen, wie z. B. Referenzsprache, visuelle Informationen, Richtungsinformationen oder Sprecheraktivität. Ein typisches SC-TSE-System besteht aus einem speaker embedder network und einem speaker separator network. Das speaker embedder network generiert zielsprecherspezifische diskriminative Merkmale aus den Zusatzinformationen, welche das speaker separator network bei der Extraktion des Zielsprechers aus dem Mischsignal leiten. Ziel dieser Arbeit ist die Entwicklung und Evaluierung neuartiger DNN-basierter Architekturen zur Verbesserung der Zuverlässigkeit, Effizienz und Robustheit einkanaliger SC-TSE-Algorithmen unter Verwendung von Referenzsprache als Zusatzinformation.

Zunächst schlagen wir drei neuartige Varianten von Long Short-Term Memory (LSTM)-Zellen für die Zielsprecherextraktion im Zeit-Frequenz-Bereich vor. Diese angepassten LSTM-Zellen sind speziell für die SC-TSE-Aufgabe konzipiert, indem sie optimieren, wie Zielsprecherinformationen innerhalb der LSTM-Zellen gespeichert und aktualisiert werden. Die erste vorgeschlagene Variante passt lediglich das Forget-Gate an, wodurch eine selektive Beibehaltung der Zielsprecherinformationen ermöglicht wird, während Informationen aus anderen Quellen im Mischsignal ignoriert werden. Die zweite vorgeschlagene Variante erweitert die erste, indem sowohl das Input- als auch das Forget-Gate angepasst werden, um den Aktualisierungsmechanismus des Zellzustands zu verbessern und die Beibehaltung zielsprecherspezifischer Merkmale zu verstärken. Die dritte vorgeschlagene Variante führt ein zusätzliches Auxiliary-Modulation-Gate innerhalb der LSTM-Zelle ein, das dazu dient, sowohl langfristige als auch kurzfristige zielsprecherspezifische Merkmalsdiskriminierung dynamisch zu erlernen. Experimentelle Ergebnisse mit verschiedenen Mischsignaltypen zeigen, dass alle vorgeschlagenen Varianten der LSTM-Zellen den Standard-LSTM-Zellen sowohl im unidirektionalen als auch im bidirektionalen Modus überlegen sind. Die besten Ergebnisse werden mit den Auxiliary-Gated-LSTM-Zellen erzielt, die eine Verbesserung des scale-invariant signal-to-distortion ratio (SI-SDR) von bis zu 1,14 dB (unidirektionaler Modus) bzw. 1,09 dB (bidirektionaler Modus) gegenüber den Standard-LSTM-Zellen erreichen.

Zweitens schlagen wir zwei auf Conformern basierende Architekturen für die Zielsprecherextraktion im Zeitbereich vor. Die erste vorgeschlagene Architektur,

Conformer-FFN, verwendet Stapel von Conformer- und externen Feed-Forward-Blöcken, mit dem Ziel, sowohl lokale als auch globale Kontextmerkmale mithilfe von Conformer-Blöcken zu nutzen, während gleichzeitig die Gesamtanzahl der Parameter durch externe Feed-Forward-Blöcke reduziert wird. Die zweite vorgeschlagene Architektur, TCN-Conformer, verwendet Stapel von Temporal Convolutional Network (TCN)- und Conformer-Blöcken, mit dem Ziel, die besten lokalen Kontextmerkmale durch TCN-Blöcke zu extrahieren und anschließend sowohl lokale als auch globale Kontextmerkmale mithilfe von Conformer-Blöcken zu nutzen. Experimentelle Ergebnisse mit verschiedenen Mischsignaltypen zeigen, dass das vorgeschlagene TCN-Conformer-System sowohl dem TCN-basierten Basissystem als auch dem vorgeschlagenen Conformer-FFN-System überlegen ist. Die besten Ergebnisse werden mit vier Stapeln aus TCN- und Conformer-Blöcken erzielt, was eine Verbesserung des SI-SDR von bis zu 2,64 dB gegenüber dem TCN-basierten Baseline und bis zu 3,44 dB gegenüber dem Conformer-FFN-System ergibt. Um das vorgeschlagene TCN-Conformer-System besser für die Echtzeit-Zielsprecherextraktion geeignet zu machen, ersetzen wir die traditionelle Multi-Head Self-Attention (MHSA) in jedem Conformer-Block des speaker separator network durch eine lineare MHSA. Experimentelle Ergebnisse zeigen, dass das TCN-Conformer-System mit linearer MHSA dem System mit traditioneller MHSA überlegen ist, während es gleichzeitig eine signifikante Reduktion der Rechenkosten und des Echtzeitfaktors erreicht. Darüber hinaus zeigen wir, dass durch Multi-Condition-Training die Robustheit gegen Hintergrundgeräusche, Nachhall und intrinsische Variabilität (z. B. Emotionen) in der Referenzsprache des Zielsprechers verbessert werden kann.

Drittens evaluieren wir die Leistung von zwei SC-TSE-Algorithmen subjektiv durch Hörtests mit normalhörenden (NH) und hörgeschädigten (HI) Testpersonen: ein Algorithmus, der die Zielsprecherextraktion mithilfe einer reellwertigen Maske im Zeit-Frequenz-Bereich durchführt (Algo-1), und ein Algorithmus, der die Zielsprecherextraktion im Zeitbereich unter Verwendung der vorgeschlagenen TCN-Conformer-Architektur durchführt (Algo-2). Diese Algorithmen wurden für herausfordernde akustische Szenarien mit bis zu zwei Störsprechern unter Anwendung von drei subjektiven Bewertungsmethoden evaluiert: Paarvergleich, Sprachverständlichkeitsschwellen (SRTs) und kategorial skalierter wahrgenommener Höraufwand. Die Ergebnisse mit fünfzehn NH- und fünfzehn HI-Testpersonen zeigen, dass Algo-2 den Höraufwand signifikant reduziert, die Sprachverständlichkeit verbessert und sowohl gegenüber den unbearbeiteten Mischsignalen als auch gegenüber Algo-1 bevorzugt wird. Darüber hinaus profitieren HI-Testpersonen stärker als NH-Testpersonen, z. B. in Bezug auf den Höraufwand mit einer Reduktion von 7–8 Einheiten (ESCU) im Vergleich zu 4–5 Einheiten bei NH-Testpersonen. Für HI-Testpersonen mit symmetrischem leichtem bis mäßigem Hörverlust legen die Ergebnisse zudem nahe, dass eine Hörverlustkompensation nicht erforderlich ist, um vom Algorithmus zu profitieren.

GLOSSARY

Acronyms and Abbreviations

AAD auditory attention decoding
ASR automatic speech recognition

BSS blind source separation

CE cross-entropy loss

DNN deep neural network
DOA direction of arrival
DRR direct-to-reverberant ratio

GMM Gaussian mixture model

HI hearing-impaired

ICA independent component analysis
IVA independent vector analysis
IVE independent vector extraction

LSTM long short-term memory

MACS multiply-accumulate operations per second
MCT multi-condition training
MHSA multi-head self-attention

NH normal-hearing

PESQ perceptual evaluation of speech quality

RIR room impulse response
RTF real-time factor

SC-TSE speaker-conditioned target speaker extraction
SI-SDR scale-invariant signal-to-distortion ratio
SNR signal-to-noise ratio
SPL sound pressure level
SRT speech recognition threshold
STFT short-time Fourier transform
STOI short-time objective intelligibility

t-SNE t-distributed stochastic neighbor embedding
TCN temporal convolutional network

UBM Universal Background Model

WSJ Wall Street Journal

Operators

$\cdot$	Dot product
$\odot$	Point-wise multiplication
σ	Sigmoid activation function
softmax	Softmax operation
softmax_q	Row-wise softmax operation
softmax_k	Column-wise softmax operation
tanh	Hyperbolic tangent activation

Fixed Symbols

j	Index of the target speaker
k	Frequency index
l	Time-frame index
n	Discrete-time index
D	Speaker embedding dimension
I	Number of speakers in the mixture
$a_j(n)$	Reference speech of the target speaker
$x_i(n)$	i-th speaker signal
$x_j(n)$	Target speaker signal
$\hat{x}_j(n)$	Estimated target-speaker signal
$v(n)$	Background noise
$\hat{v}(n)$	Estimated background noise
$y(n)$	Mixture signal
$M_j(k,l)$	Real-valued mask of target speaker
$X_i(k,l)$	Speech component corresponding to the i-th speaker in the STFT domain
$X_j(k,l)$	Speech component corresponding to the target speaker in the STFT domain
$\hat{X}_j(k,l)$	Estimated target speech component in the STFT domain
$V(k,l)$	Noise component in the STFT domain
$Y(k,l)$	Mixture signal in the STFT domain
$\mathbf{e}_j$	Target speaker embedding vector
$\mathbf{r}$	Transformed mixture feature representation
$\mathbf{a}_t$	Auxiliary-modulation gate

$\mathbf{b}_a$	Bias vector for auxiliary-modulation gate
$\mathbf{b}_c$	Bias vector for cell update
$\mathbf{b}_f$	Bias vector for forget gate
$\mathbf{b}_i$	Bias vector for input gate
$\mathbf{b}_o$	Bias vector for output gate
$\mathbf{b}_{ef}$	Bias vector for customized forget gate (F)
$\mathbf{b}_{ei}$	Bias vector for customized input gate (F+I)
$\mathbf{c}_t$	Current LSTM cell state
$\mathbf{c}_{t-1}$	Previous LSTM cell state
$\mathbf{f}_t$	Forget gate of LSTM cell
$\mathbf{h}_t$	Current hidden state
$\mathbf{h}_{t-1}$	Previous hidden state
$\mathbf{i}_t$	Input gate of LSTM cell
$\mathbf{o}_t$	Output gate of LSTM cell
$\mathbf{W}_a$	Weight matrix for auxiliary-modulation gate
$\mathbf{W}_c$	Weight matrix for cell update
$\mathbf{W}_f$	Weight matrix for forget gate
$\mathbf{W}_i$	Weight matrix for input gate
$\mathbf{W}_o$	Weight matrix for output gate
$\mathbf{W}_{ef}$	Weight matrix for customized forget gate (F)
$\mathbf{W}_{ei}$	Weight matrix for customized input gate (F+I)
$\mathbf{K}$	Key matrix for self-attention mechanism
$\mathbf{Q}$	Query matrix for self-attention mechanism
$\mathbf{V}$	Value matrix (self-attention)
$Att(\mathbf{Q}, \mathbf{K}, \mathbf{V})$	Traditional attention matrix
$Att_{\text{lin}}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$	Linear attention matrix
$\boldsymbol{\phi}(\cdot)$	LSTM block function
$\boldsymbol{\phi}^{\text{emb}}$	Speaker embedder network (time-domain SC-TSE)
$\boldsymbol{\phi}^{\text{sep}}$	Speaker separator network (time-domain SC-TSE)
$\mathcal{F}$	SC-TSE model
$\mathcal{F}_{bss}$	BSS model
$\mathcal{F}_{se}$	Speech enhancement model
$\mathcal{L}_{\text{AUD}}$	Auditory-inspired loss function
$\mathcal{L}_{\text{CE}}$	Cross-entropy loss function
$\mathcal{L}_{\text{MSE}}$	Mean-square error loss function
$\mathcal{L}_{\text{MS-SI-SDR}}$	Multi-scale SI-SDR loss function
$\mathcal{L}_{\text{SI-SDR}}$	SI-SDR loss function

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

The cocktail-party problem [1] represents a complex acoustic scenario in which a listener attempts to follow the conversation of a target speaker in the presence of multiple interfering speakers and background noise. Despite these challenges, humans are very good at selectively attending to the target speaker while suppressing all other sound sources, a process governed by the selective attention mechanism [2]. The human brain accomplishes this by integrating multiple perceptual cues, including spatial cues (identifying the direction of the target speaker), spectral cues (distinguishing vocal tone and frequency), visual cues (observing lip movements and facial expressions), and semantic cues (understanding conversational context). While humans perform this task easily, most speech processing algorithms struggle under such conditions. Although the exact mechanisms behind selective attention are not yet fully understood, one of the primary goals of speech processing research is to develop algorithms that can mimic these abilities and effectively extract the target speaker from a mixture. Over the past few decades, significant research efforts have been dedicated to addressing this challenge.

Traditionally, extracting a single speaker from a mixture has been approached as a speaker separation task [3], [4], which aims to extract all individual sources. However, in many practical applications, such as hearing aids, assistive listening devices, teleconferencing, broadcasting and live streaming, and automatic speech recognition (ASR), it is often not required to separate all sources; instead, extracting only the target speaker suffices. Typically, this can be achieved by first performing speaker separation using a blind source separation (BSS) algorithm [5]–[12] and then selecting the extracted source corresponding to the target speaker. However, BSS algorithms face a key challenge: they typically require knowledge or an estimate of the number of sources present in the mixture, which is not trivial in practice.

Recent advances in deep neural networks (DNNs) have enabled a more direct solution, known as speaker-conditioned target speaker extraction (SC-TSE) [13]. SC-TSE algorithms treat the target speaker extraction task as a binary classification problem, where the target speaker is assigned to the positive class and all other sources to the negative class. In general, these algorithms utilize auxiliary information of the target speaker to distinguish the target speaker from other sources in the mixture [14]–[51]. Such auxiliary information may include reference speech (i.e., a short pre-recorded segment of the target speaker) [14]–[25], video signal [26]–[36], spatial information [37]–[41], or speech activity [42] of the target speaker.

Various DNN-based systems have been proposed for SC-TSE, demonstrating promising performance. However, several challenges remain, including mitigating the mismatch between training and test conditions, ensuring robustness against unseen noise environments, and enabling real-time processing. Moreover, SC-TSE algorithms are typically evaluated using objective performance metrics, which do not fully capture human perception of speech quality and intelligibility. Therefore, the main objective of this thesis is to **develop and evaluate (both objectively and subjectively) novel DNN-based architectures leveraging reference speech of the target speaker as auxiliary information to enhance the reliability, efficiency, and robustness of single-channel SC-TSE algorithms.**

The remainder of this chapter is structured as follows. In **Section 1.1** we define the target speaker extraction problem. In **Section 1.2** we discuss the relationship between target speaker extraction and other speech processing tasks. In **Section 1.3** we discuss several factors that influence the performance of target speaker extraction. In **Section 1.4** we discuss different forms of auxiliary information that can be used for target speaker extraction with SC-TSE algorithms. Finally, in **Section 1.5** we present the outline of the thesis and highlight our main contributions.

1.1 Problem definition

We consider a scenario in which a single microphone aims to record a specific target speaker, but also picks up voices from other (interfering) speakers along with background noise. The recorded mixture signal $y(n)$ at the microphone, with n discrete time index, can be expressed as

$$y(n) = \sum_{i=1}^{I} x_i(n) + v(n), \tag{1.1}$$

where $x_i(n)$ denotes the speech signal corresponding to the i-th speaker. I denotes the number of speakers and $v(n)$ denotes the background noise signal. The j-th speaker is assumed to be the target speaker, with $1 \leq j \leq I$.

In this thesis, we address this challenge by directly extracting the target speaker while suppressing all other speakers and background noise by employing an SC-TSE algorithm. To characterize the target speaker, we assume access to single-channel reference speech of the target speaker as auxiliary information, denoted as $a_j(n)$. The goal of an SC-TSE algorithm (see Fig. 1.1) is to estimate the target speaker, $x_j(n)$, given the recorded mixture signal $y(n)$ and the reference speech of the target speaker $a_j(n)$. We define this processing as

$$\hat{x}_j(n) = \mathcal{F}\left(y(n), a_j(n)\right), \tag{1.2}$$

where $\hat{x}_j(n)$ is the estimated target speaker and $\mathcal{F}$ represents the model for a general SC-TSE algorithm.

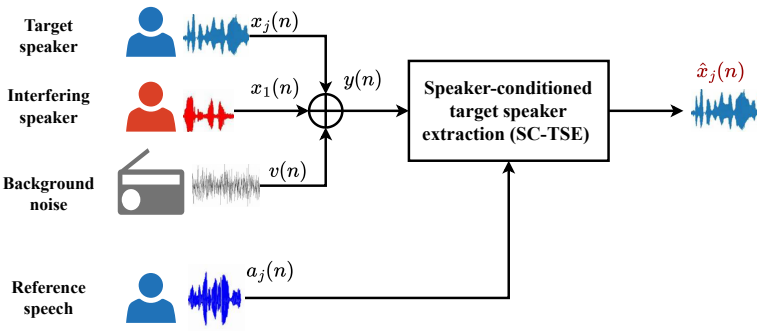

Fig. 1.1: A speaker-conditioned algorithm for target speaker extraction utilizing auxiliary information about the target speaker. For simplicity, the figure illustrates the scenario of a mixture containing the target speaker, one interfering speaker, and background noise.

1.2 Target speaker extraction and its relation to other tasks

Target speaker extraction is closely related to BSS and speech enhancement. While both target speaker extraction and speech enhancement aim to remove undesired sources from the mixture, BSS aims to estimate all individual sources. Due to this relationship, many approaches developed for one task can be adapted for the other task. The key difference lies in the availability of auxiliary information. Target speaker extraction utilizes this information to guide the separation process, whereas BSS and speech enhancement operate without any additional guidance. Beyond BSS and speech enhancement, target speaker extraction also shares similarities with other speech processing tasks such as speaker diarization and speaker adaptation. This section explores these relationships in more detail: Sections 1.2.1 and 1.2.2 discuss the relationship between target speaker extraction and BSS and between target speaker extraction and speech enhancement, respectively. In Section 1.2.3, we briefly discuss how target speaker extraction relates to speaker diarization and speaker adaptation.

1.2.1 *Relation to BSS*

BSS aims to estimate all individual sources from the mixture without relying on any auxiliary information. This can be expressed as

$$\{\hat{x}_1(n), \hat{x}_2(n), ..., \hat{x}_I(n), \hat{v}(n)\} = \mathcal{F}_{\text{bss}}\left(y(n)\right), \tag{1.3}$$

where $\hat{x}_i(n)$ and $\hat{v}(n)$ denote the estimated speech sources and background noise, and $\mathcal{F}_{\text{bss}}$ represents the BSS model.

To perform target speaker extraction using BSS, two steps are required (see Fig. 1.2). First, each source of the mixture is estimated using a BSS algorithm, and

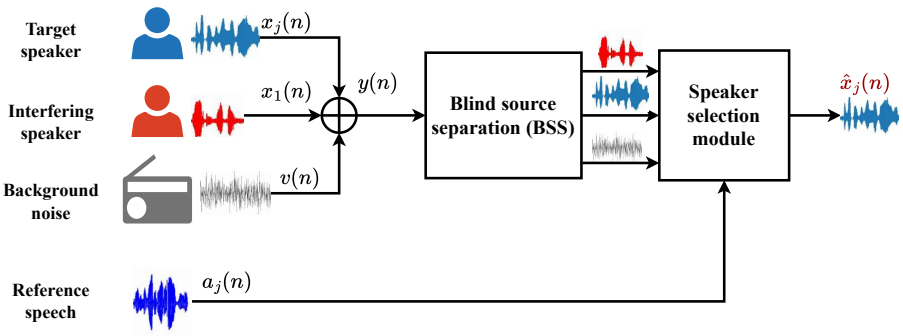

Fig. 1.2: Target speaker extraction using a blind source separation algorithm (referred to as two-step method).

then a speaker selection module is used to select the source corresponding to the target speaker. For instance, a speaker identification system [52], [53] is used to compare each estimated source against the reference speech of the target speaker and select the most likely target speaker [54]. The speaker selection step aims to minimize intra-speaker distance while maximizing inter-speaker separability. The selected target speaker ($\hat{x}_j(n)$) from the speaker selection module can be expressed as:

$$\hat{x}_j(n) = TSS\left((\hat{x}_1(n), ..., \hat{x}_I(n), \hat{v}(n)), a_j(n)\right), \tag{1.4}$$

where $TSS\left(\cdot\right)$ represents the speaker selection module.

1.2.2 *Relation to speech enhancement*

Speech enhancement [55]–[61] aims to enhance the target speaker from the mixture while simultaneously removing all undesired sources, without relying on any auxiliary information (see Fig. 1.3). This can be expressed as

$$\hat{x}_j(n) = \mathcal{F}_{\text{se}}\left(y(n)\right), \tag{1.5}$$

where $\hat{x}_j(n)$ denotes the estimated target speaker and $\mathcal{F}_{\text{se}}$ represents the speech enhancement model.

Speech enhancement algorithms work under the assumption that the target speaker and undesired sources have distinct spectro-temporal characteristics, enabling effective discrimination between them.

Early research on target speaker extraction framed target speaker extraction as a speech enhancement problem, either by extracting the dominant speaker [62] or by using speaker-specific DNNs [63], [64]. However, these approaches often fail when interfering speakers are equally loud or louder than the target speaker. Additionally, they cannot generalize to unseen target or interfering speakers, limiting their appli-

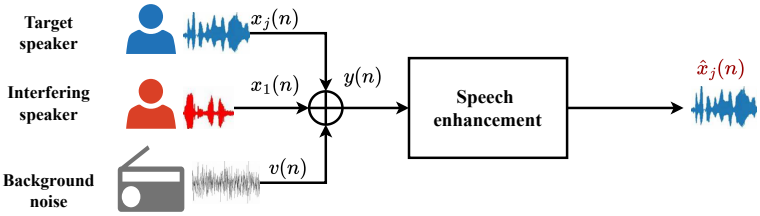

Fig. 1.3: Speech enhancement algorithm to estimate the desired target speaker while simultaneously removing all undesired sources from the mixture.

cability in real-world scenarios. In contrast, SC-TSE algorithms extend conventional speech enhancement by introducing speaker awareness, allowing the extraction of a specific target speaker even in the presence of interfering speakers.

1.2.3 *Relation to other speech processing tasks*

Target speaker extraction is also related to other speech processing tasks, such as speaker diarization and speaker adaptation in ASR. Speaker diarization aims to determine "who speaks when" [65]–[71], typically by identifying the number of speakers in the mixture and their respective speaking time intervals. While this is traditionally performed using clustering-based algorithms [65], [66], Speaker-conditioned techniques can also be utilized for diarization if auxiliary information for each speaker is available. For instance, in [67], a DNN-based algorithm is proposed that utilizes speaker embeddings to directly estimate the voice activity of each target speaker at every time frame. Similarly, in [68] an algorithm that jointly estimates speaker embeddings and detects speaker activity frame-by-frame, without requiring prior knowledge of the number of speakers is proposed, making it well-suited for scenarios with a varying number of speakers. In [69], both diarization and speaker separation are jointly performed, conditioned on the speaker embedding of each speaker. Even when auxiliary information for each speaker is not initially available, as is often the case in diarization, it could potentially be derived from single-speaker segments identified through an initial diarization [67], and then iteratively be used to refine diarization performance. Furthermore, in [71] a speaker-conditioned diarization approach focused on computational efficiency is proposed, where speaker attractors are estimated in parallel, which employs an iterative refinement process in which both attractors and diarization outputs are repeatedly updated for improved alignment. However, despite its feasibility, speaker-conditioned diarization algorithms are generally more complex and computationally demanding than traditional diarization algorithms [70].

Similarly, speaker adaptation in ASR also shares conceptual similarities with target speaker extraction. In speaker adaptation, the algorithm adjusts to a specific speaker using auxiliary information, typically in the form of reference speech [72], [73], making it closely related to SC-TSE. However, the key difference between

speaker adaptation and SC-TSE lies in their degree of dependency on the reference speech. Speaker adaptation aims to make minimal adjustments to the output of the algorithm based on the reference speech, allowing the algorithm to function reasonably well even without adaptation. In contrast, SC-TSE algorithms highly depend on the reference speech.

1.3 Factors influencing target speaker extraction

The performance of target speaker extraction is affected by various factors, including the amount of background noise and reverberation, the percentage of overlap between speakers, speaker-specific characteristics, inactive target speaker, and mismatch between training and testing conditions. These factors will be discussed in more detail in the following subsections.

1.3.1 *Background noise and reverberation*

Background noise and reverberation are two prominent acoustic factors that significantly affect the quality and intelligibility of the target speaker in real-world environments. Background noise is an inevitable part of our daily lives, present in nearly every environment, from busy streets and restaurants to offices and living rooms. It comes in various forms, often interfering with the target speaker. Background noise can be broadly categorized based on its temporal characteristics. Stationary noise, such as the hum of an air conditioner or a fan, remains relatively constant over time. In contrast, non-stationary noise, such as a passing siren or a honking car, changes dynamically over time, making it more challenging to remove from the mixture. Additionally, noise can be classified based on its spectral characteristics. Narrowband noise is concentrated within a limited frequency range, while broadband noise spreads across the entire frequency range. The level of background noise varies across environments and is typically measured in decibels (dB) of sound pressure level (SPL). SPL is a measure that quantifies acoustic pressure relative to the threshold of human hearing. In quiet environments, such as hospitals or classrooms, noise levels typically range from 50 dB to 55 dB SPL, whereas in loud environments, such as airplanes or trains, levels can reach 70 dB to 75 dB SPL. As the noise level increases, the signal-to-noise ratio (SNR) reduces, which makes speaker extraction more challenging.

In addition to background noise, reverberation further distorts the target speaker by altering its temporal and spectral characteristics [74], [75]. In an indoor environment, the speech from the target speaker is repeatedly reflected off walls and other surrounding objects, producing multiple delayed and decaying copies of the original target speaker signal. This phenomenon, known as reverberation, is characterized by the room impulse response (RIR). The RIR can be decomposed into three components: the direct path, the component arriving via the shortest path; early reflections, which arrive within the first 50 milliseconds following the direct path; and late reflections, consisting of multiple weak reflections that arrive more than 50 milliseconds after the direct path. Late reflections are the primary cause

of reduced intelligibility of the target speaker. Several acoustic properties of the environment can be derived directly from the RIR. One such property is the reverberation time (T_{60}), which measures the time required for the late reverberation to decay by 60 dB relative to the direct sound level [74]. Another important property is the direct-to-reverberant ratio (DRR), which measures the ratio of energy in the direct path to that in the reverberant components [76].

1.3.2 *Percentage of overlap between speakers*

The percentage of overlap between the target speaker and interfering speakers is another factor influencing the performance of both speech separation [77] and target speaker extraction [42]. This overlap quantifies the duration during which multiple speakers are active simultaneously in the mixture. Higher overlap increases the complexity of the task due to higher temporal and spectral masking of the target speaker, while lower overlap results in less masking and easier target speaker extraction. When the overlap is low ($0-20\%$), extracting the target speaker is relatively straightforward, as each speaker tends to occupy distinct time-frequency bins. As the overlap increases to a moderate range ($20-50\%$), extraction becomes more challenging due to increased time-frequency overlap, resulting in greater spectro-temporal ambiguity between speakers. At extreme range ($80-100\%$), this ambiguity becomes severe, making it highly difficult to distinguish the target speaker, even for human listeners.

1.3.3 *Speaker-specific characteristics*

Speaker-specific characteristics are another factor that influences the performance of both speech separation [78], [79] and target speaker extraction [80]–[82]. While in [78] the influence of speaker gender is primarily investigated, in [79] the investigation is extended to speaker-specific characteristics, regardless of speaker gender, in the context of speech separation. A key finding of this research is that differences in fundamental frequency between the speakers significantly influence separation performance, whereas the performance is hardly influenced by differences in vocal tract length. Similarly, in [80] the influence of speaker emotion on both tasks is investigated, while in [81], [82] the influence of speaker gender and similarity in speaker-specific characteristics is investigated in the context of target speaker extraction. These studies also highlighted that factors such as speaker emotion, fundamental frequency, and speaking style significantly influence extraction performance.

1.3.4 *Domain mismatch*

Recent advancements in speech separation and target speaker extraction primarily rely on data-driven algorithms. These algorithms use DNNs to learn the underlying structure of the data during training, which is then applied during testing. However, their performance significantly degrades when there is a mismatch between

the training and testing conditions [20], [83]. Another major challenge arises from the fact that these algorithms are typically trained on artificially simulated data, which does not fully capture the complexities of real-world environments. When tested in a real-world environment, various factors, such as the natural percentage of overlap between speakers in the mixture, realistic room impulse responses, and intrinsic variability, can lead to degraded performance. Additionally, similar speech characteristics of different speakers [81], [82], and language differences [84] further complicate the target speaker extraction task.

1.3.5 *Inactive target speaker*

Most target speaker extraction algorithms work under the assumption that the target speaker is always present and active within the mixture. However, in realistic scenarios, this assumption does not always hold, as the presence of the target speaker is often unpredictable or unknown in advance [85]–[87]. Ideally, in such scenarios, the algorithm should produce a zero-valued output when the target speaker is inactive in the mixture. However, many existing algorithms still generate an output signal in this case, which may either be a distorted signal or an unintended extraction of an interfering speaker. This unintended behavior reduces the reliability of the algorithm and poses challenges to its practical application.

This thesis examines the influence of several factors on the performance of SC-TSE algorithms for different DNN-based architectures. Throughout this thesis, we assume that the mixture signals are anechoic and the overlap between the target and interfering speakers is high (fully overlapped). Chapters 3 and 4 specifically focus on the impact of background noise and domain mismatch in the input mixture. Appendix A further explores the effects of noise, reverberation, speech characteristics, and domain mismatch in the auxiliary information.

1.4 Auxiliary information

Although in this thesis we focus on using reference speech as auxiliary information for target speaker extraction, other forms of auxiliary information can be utilized. Sections 1.4.1 and 1.4.2 discuss the most common types of auxiliary information: reference speech and visual information, while Section 1.4.3 discusses the speaker activity-based auxiliary information. Although this thesis focuses on single-channel SC-TSE, multi-microphone algorithms utilizing spatial information have been proposed, which is discussed in Section 1.4.4. Section 1.4.5 discusses the emerging use of brain signals as auxiliary information.

1.4.1 *Reference speech*

The use of reference speech as auxiliary information has gained significant attention [14], particularly as a solution to the generalization challenge faced by speaker-

specific target speaker extraction algorithms, which require large amounts of target speaker data [63]. Reference speech is a short recording of the target speaker, typically only a couple of seconds. It is one of the simplest and most practical forms of auxiliary information and has been widely used in SC-TSE algorithms [14]–[25], [44]. Since it can be easily obtained without additional hardware, such as cameras, it is well suited for personalized applications, where users can pre-record a speech sample to facilitate speaker extraction. Additionally, in long recordings (e.g., of a meeting) single-speaker segments can also be utilized as reference speech.

Although reference speech is the simplest and most widely used form of auxiliary information, its performance can be limited due to inter- and intra-speaker variability. For instance, the voice characteristics of family members may be highly similar [88], making discrimination difficult [81], [82]. Moreover, the speech of the target speaker may vary due to factors such as emotions [80], health, or aging [89], posing additional challenges for the reliable extraction of the target speaker.

1.4.2 *Visual information*

Visual information has been widely used in speech separation [26]–[30] and target speaker extraction [31]–[36]. The research in [90] suggests that visual information can help humans to focus on the specific speaker, making them a natural choice for auxiliary information in SC-TSE algorithms. In visual information-based SC-TSE algorithms, the auxiliary information is derived from video signals, typically utilizing the face of the target speaker, lip movements, or both. These algorithms follow a similar approach to SC-TSE algorithms using reference speech. However, in real-world scenarios, partial occlusions of the face of the target speaker can limit the effectiveness of visual information-based SC-TSE algorithms [91]. To address these challenges, multimodal approaches that combine both reference speech and visual information have been explored [92]–[96]. Additionally, in [97] it has been investigated to use still images as auxiliary information, assuming that facial features alone can provide the discriminative features between the target and interfering speakers. While this image-based SC-TSE algorithm has shown modest performance, it becomes particularly effective in scenarios where speakers have distinct visual features, such as different genders.

1.4.3 *Speaker activity*

In [42] it has been proposed to utilize speaker activity as auxiliary information in SC-TSE algorithms. Several visual information-based SC-TSE algorithms use only the lip movement [31], [33], [34] extracted from the video signal of the target speaker, indicating that speaker activity plays a crucial role in discriminating the target speaker from the interfering speakers. Since lip movements provide strong information for speech activity, this hypothesis has been explored in [42]. When using the same architecture, the SC-TSE algorithm utilizing the speaker activity information has shown similar performance to the version using reference speech as auxiliary information.

1.4.4 *Spatial information*

For multi-microphone SC-TSE algorithms, spatial information (e.g., the location or direction of arrival (DOA) of the target speaker relative to the recording device) is another type of auxiliary information that can be used. Spatial information can be estimated, e.g., from a video or from a previous recording of the target speaker in a known position. Several SC-TSE algorithms [37]–[41] have exploited the spatial information for target speaker extraction and have shown promising results. Furthermore, some of the algorithms [44], [45] have also exploited the combinations of different types of auxiliary information, such as reference speech and spatial information in [44], and reference speech, visual information, and spatial information in [45] to improve the performance of target speaker extraction in challenging real-world environments.

1.4.5 *Brain signals*

Research on auditory attention has shown a strong correlation between the attended speech of a listener and their neural responses, which can be measured using electroencephalography (EEG) [98], [99]. This hypothesis has recently been explored in some SC-TSE algorithms [49]–[51], utilizing EEG-derived attention as auxiliary information. These algorithms rely on auditory attention decoding (AAD) to determine which speaker in the mixture aligns with the neural activity of the listener, enabling a listener-guided approach for speaker extraction. As a result, EEG-driven SC-TSE algorithms hold great potential for hearing aid applications and advanced auditory processing systems. Although EEG-driven target speaker extraction is still in its early stages, it is rapidly gaining research interest due to its unique ability to extract the target speaker based on listener intent.

1.5 Outline of the thesis and main contributions

The main objective of this thesis is to **develop and evaluate novel DNN-based architectures leveraging reference speech of the target speaker to enhance the reliability, efficiency and robustness of single-channel SC-TSE algorithms.** The first focus is to improve the speaker extraction performance and the real-time capability of SC-TSE algorithms using long short-term memory (LSTM) and conformer-based architectures. The second focus is to subjectively evaluate SC-TSE algorithms through listening tests with normal-hearing (NH) and hearing-impaired (HI) listeners to assess their real-world applicability.

The main contributions of this thesis are threefold. As a first contribution, **we propose three novel LSTM cell variants for the speaker separator network of an SC-TSE algorithm operating in the time-frequency domain. These customized LSTM cells are specifically designed for the SC-TSE task**, more in particular to either selectively retain information relevant to the target speaker, simultaneously retain and update target speaker-related information,

or effectively learn both short-term and long-term discriminative features. Experimental results across various mixture types show that all proposed LSTM variants outperform an LSTM-based baseline system, with the best results achieved by the LSTM variant optimized for learning both short-term and long-term discriminative features. As a second contribution, **we propose two conformer-based architectures for the speaker separator network of an SC-TSE algorithm operating in the time domain**. The first architecture uses stacks of conformer and external feed-forward blocks (Conformer-FFN), while the second architecture uses stacks of temporal convolutional network (TCN) and conformer blocks (TCN-Conformer). These architectures are designed to effectively capture both local and global context features. Experimental results across various mixture types show that the proposed TCN-Conformer system significantly improves the target speaker extraction performance compared to the proposed Conformer-FFN system and a TCN-based baseline system. In addition, we propose to replace traditional multi-head self-attention (MHSA) in the TCN-Conformer architecture with linear MHSA, significantly reducing computational complexity while maintaining performance. As a third contribution, **we subjectively evaluate two SC-TSE algorithms using listening tests with NH and HI listeners**: an algorithm performing target speaker extraction using a real-valued mask in the time-frequency domain (Algo-1) and an algorithm performing target speaker extraction in the time domain using the proposed TCN-Conformer architecture (Algo-2). These algorithms were evaluated in challenging acoustic scenarios with up to two interfering speakers using three methods: paired comparison, speech recognition thresholds, and categorically scaled perceived listening effort. Results show that Algo-2 significantly reduces listening effort, improves speech intelligibility, and is preferred by the listeners compared to both the unprocessed mixture and Algo-1. Furthermore, results also suggest that HI listeners benefit more than NH listeners, especially for the female interfering speakers, and hearing loss compensation is not necessary to achieve algorithmic benefits.

In the remainder of this section, we provide a chapter-by-chapter overview of this thesis, describing the content and contribution of each chapter. A schematic overview of the thesis is depicted in Fig. 1.4.

In **Chapter 2**, we provide an overview of state-of-the-art target speaker extraction algorithms, focusing on DNN-based approaches. This chapter covers key aspects of SC-TSE algorithms, such as methods for conditioning the DNN about the target speaker, methods for computing speaker embeddings from auxiliary information, operating domains of SC-TSE algorithms, loss functions, and existing DNN-based architectures. This chapter provides an overview of classical approaches for speech separation and target speaker extraction, and discusses performance metrics to evaluate the performance of SC-TSE algorithms.

In **Chapter 3**, we consider an LSTM-based algorithm performing SC-TSE in the time-frequency domain [15] and propose three novel variants of LSTM cells for the speaker separator network, which are customized specifically for the SC-TSE task. The first proposed variant focuses on customizing only the forget gate, aiming at

retaining only target speaker information in the cell state, while the second variant focuses on both the forget and input gates, aiming at a more effective resetting of the cell state by retaining and updating only target speaker information and simultaneously disregarding information from other sources. Inspired by [100], the third proposed variant introduces an additional auxiliary-modulation gate within the LSTM cell. The purpose of this auxiliary-modulation gate is to modulate the information of the forget and input gates, aiming at better learning the long-term and short-term discriminative features of the target speaker with the help of the corresponding speaker embedding. Experiments are performed on 2-speaker mixtures, 3-speaker mixtures, and noisy mixtures (containing 1, 2 or 3 speakers) simulated using the Librispeech and MUSAN datasets. Results show that all proposed variants of LSTM cells outperform standard LSTM cells in both unidirectional and bidirectional modes in terms of objective target speaker extraction measures. The best performance is obtained using the auxiliary-gated LSTM cells, which yield scale-invariant signal-to-distortion ratio (SI-SDR) improvements up to 1.14 dB (unidirectional mode), and 1.09 dB (bidirectional mode) compared to standard LSTM cells. A comparison with two-step methods (see Section 1.2.1) for 2-speaker mixtures shows SC-TSE using all proposed LSTM variants consistently outperforms two-step methods based on BSS. The content of this chapter is based on the publications [101] and [102].

As an alternative to performing SC-TSE in the time-frequency domain, in **Chapter 4** we consider SC-TSE in the time domain and propose two different conformer-based architectures. The first proposed architecture (Conformer-FFN) uses stacks of conformer and external feed-forward blocks, aiming at exploiting both local and global context features using conformer blocks, while reducing the overall number of parameters using external feed-forward blocks. The second proposed architecture (TCN-Conformer) uses stacks of TCN and conformer blocks, aiming at utilizing the best local context features using TCN blocks and then exploiting both local and global context features using conformer blocks. Experimental results on 2-speaker mixtures, 3-speaker mixtures, and noisy 2-speaker mixtures simulated using the WSJ0 and WHAM datasets show that the proposed TCN-Conformer system outperforms the TCN-based baseline system in [17] and the proposed Conformer-FFN system. The best performance is obtained with four stacks of the TCN and conformer blocks, which yield SI-SDR improvements up to 2.64 dB over the TCN-based baseline and up to 3.44 dB over the Conformer-FFN system. In addition, we perform two modifications to the proposed TCN-Conformer system to make it more suitable for real-time target speaker extraction. First, we replace the traditional MHSA in each conformer block of the speaker separator network with linear MHSA. Second, we reduce the overall number of parameters by factors of 2, 4, and 8. Experimental results show that the TCN-Conformer system using linear MHSA outperforms the TCN-Conformer system using traditional MHSA for all considered network sizes while achieving a substantial reduction in computational cost and real-time factor. The content of this chapter is based on the publications [103] and [104].

While in the previous chapters the performance of SC-TSE algorithms is only evaluated using objective measures, in **Chapter 5** we subjectively evaluate the perfor-

mance by performing listening tests with NH and HI listeners (with and without linear hearing loss compensation). We have considered two different SC-TSE algorithms. Algo-1 employs a ResNet-GRU-based system to perform target speaker extraction in the time-frequency domain using a real-valued mask, while Algo-2 employs a TCN-Conformer system with traditional MHSA (we used the TCN-Conformer system proposed in chapter 4, modified by applying causal masking to MHSA at each timestep and removing the same padding from the convolution layers) to perform target speaker extraction in the time domain. Both algorithms are trained on the same dataset as in Chapter 4, but are subjectively evaluated using German matrix sentences from the Oldenburg Sentence Test (OLSA) for challenging acoustic scenarios with up to two interfering speakers. The interfering speakers, either male or female, were selected from a different matrix sentence dataset, while the target speaker was always a male speaker from OLSA. We have considered three subjective evaluation methods: paired comparison, speech recognition thresholds (SRTs), and categorically scaled perceived listening effort. The evaluation results with fifteen NH and fifteen HI listeners show that all considered methods are suitable to evaluate the performance of SC-TSE algorithms. For both NH and HI listeners, Algo-2 significantly reduces listening effort, improves speech intelligibility, and is preferred compared to the unprocessed mixtures and Algo-1, whereas Algo-1 shows no benefit compared to the unprocessed mixtures in any evaluation method. Specifically, Algo-2 reduces listening effort by 7-8 units (ESCU) for HI listeners and 4-5 units for NH listeners, and shows SRT improvements of ~ 3 dB (two male interfering speakers) and ~ 1 dB (two female interfering speakers) for HI listeners, while (for NH listeners) ~ 3 dB improvement only for male interfering speakers. Furthermore, for HI listeners with symmetric mild-to-moderate hearing loss, the results also suggest that hearing loss compensation is not necessary to obtain an algorithm benefit. The content of this chapter is based on the publications [105] and [106].

In **Chapter 6**, we summarize the main findings of the thesis and provide an outlook on potential further research.

In all previous chapters, we have assumed that the reference speech of the target speaker was clean and matched between training and testing the SC-TSE algorithms. To increase robustness against acoustic disturbances (background noise, reverberation) and intrinsic variability (emotions) in the reference speech, in **Appendix A** we explore different training possibilities for the TCN-Conformer system proposed in Chapter 4. We propose different multi-condition training (MCT) strategies by either training the entire SC-TSE system from scratch, or fine-tuning only the last layer of the speaker embedder network, or the speaker separator, or both networks. Experimental results on noisy, reverberant noisy, and emotional reference speech for 2-speaker mixtures, 3-speaker mixtures, and noisy 2-speaker mixtures simulated using the WSJ0, WHAM, and RAVDESS2mix datasets show that all utilized MCT strategies improve robustness of the SC-TSE system against mismatched reference speech, with MCT from scratch showing the best performance.

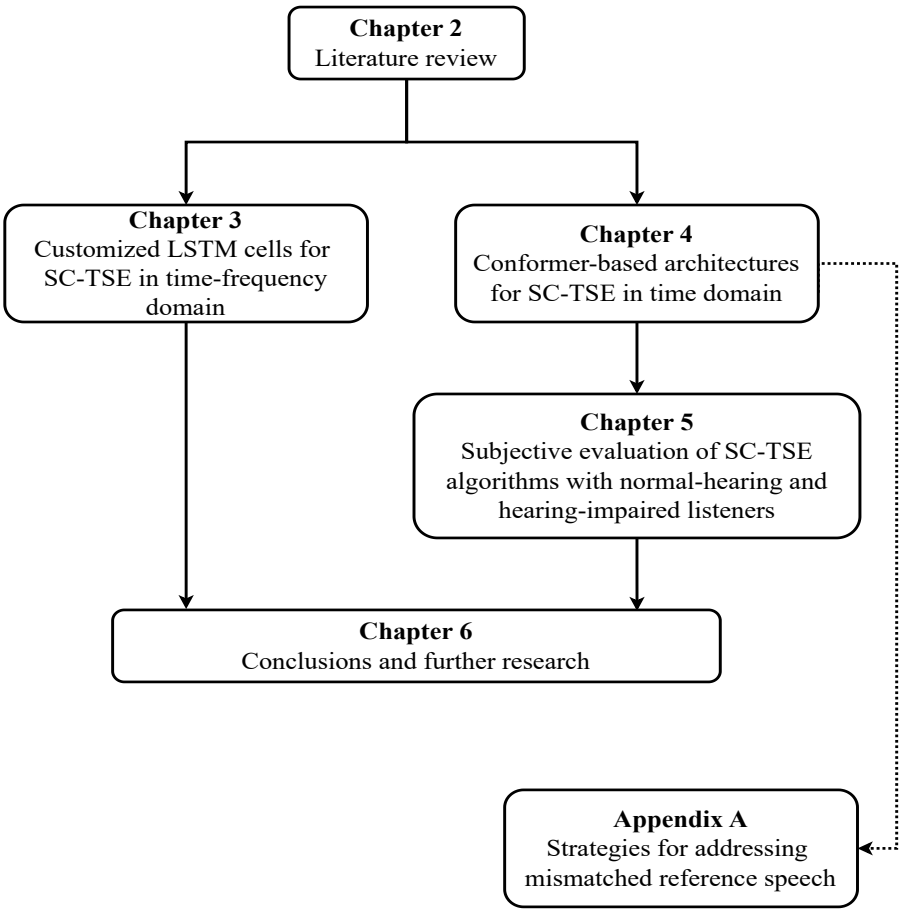

Fig. 1.4: Schematic overview of the thesis.

2

LITERATURE REVIEW

In this chapter, we provide an overview of algorithms for target speaker extraction, with a primary focus on DNN-based approaches. In Section 2.1, we briefly review some classical approaches, while in Section 2.2, we discuss DNN-based SC-TSE algorithms and their key components, such as computing speaker embeddings from auxiliary information, methods for conditioning the DNN about the target speaker, operating domains for SC-TSE algorithms (time-frequency domain, time domain), and loss functions. In Section 2.3, we discuss several existing SC-TSE algorithms, some of which are used as baseline algorithms for the thesis. Finally, in Section 2.4, we discuss the performance measures used to evaluate the speaker extraction performance.

2.1 Classical approaches

The problem of separating speakers from a mixture of overlapping speech has been extensively researched in terms of speech separation and speaker extraction. Although DNN-based approaches have recently outperformed classical approaches, several of these classical approaches are still used, either independently or in combination with DNNs to improve the performance of both speech separation and target speaker extraction. In this section, we provide an overview of some of these approaches, including beamforming (Section 2.1.1), independent component analysis (ICA) (Section 2.1.2), independent vector analysis (IVA) (Section 2.1.3), and independent vector extraction (IVE) (Section 2.1.4).

2.1.1 *Beamforming algorithms*

A typical beamforming algorithm aims to enhance the desired source (target speaker) while suppressing undesired sources (interfering speakers and background noise) in a multichannel mixture by exploiting spatial information captured by a microphone array. This is achieved by designing a spatial filter, characterized by angle and frequency, that enhances the source arriving from the target direction and suppresses sources arriving from other directions. Beamforming then applies this filter to each microphone channel and combines the outputs in such a way that on-axis components add constructively, while off-axis components cancel destruc-

tively. Beamforming algorithms can be classified into two categories: fixed (data-independent) and adaptive (data-dependent). Fixed beamformers rely on a priori knowledge of the array geometry and acoustic scenario, with parameters and models defined in advance and remain constant over time, while adaptive beamformers estimate their parameters directly from the microphone signal.

Fixed beamforming

The common example of fixed beamforming include fixed delay-and-sum, which assumes a fixed DOA for the target speaker, and fixed superdirective, which incorporates a model of the spatial coherence of the noise field, often assuming isotropic noise distributions. The simplest fixed delay-and-sum beamforming algorithm [107], [108] can be realized by assuming the DOA of the target speaker to be fixed and choosing the parameters accordingly. Fixed beamformers have been employed for both speech separation and target speaker extraction [109]–[111]. In [109], a fixed delay-and-sum beamformer is used to enhance the target speaker in a reverberant environment, while [110] extend this with a two-stage target speaker extraction framework. At first, a delay-and-sum beamformer is steered to the DOA of the target speaker, then the time-frequency representation of the beamformer outputs are used as the input to the ICA-based BSS, which estimates demixing matrices per frequency bin, with a permutation-alignment step to ensure coherent reconstruction of the target speaker. Some work [40], [112], [113] have also integrated the fixed beamforming with DNNs. For instance, in [112], a multi-beam DNN is proposed to generate multiple beam, each processed by its own DNN to estimate the mask, with a final selection of the best beam per speaker, while [113] extends the work in [112] by adding an LSTM-based beam selection module that chooses the optimal beam before the DNN. Whereas in [40], a fixed beamformer steered to the known DOA of the target speaker is used to generate an auxiliary reference signal that is then used for in SC-TSE algorithm for target speaker extraction.

Adaptive beamforming

Adaptive beamforming algorithms are in general more flexible regarding the acoustic scenario compared to fixed beamforming algorithms. However, they typically require accurate parameter estimation, such as the steering vector or spatial covariance matrices. The most widely used adaptive beamformer include the minimum power distortionless response (MPDR), minimum variance distortionless response (MVDR), and linearly constrained minimum variance (LCMV) beamformers [108], [111], [114]. MPDR beamformer minimizes the total output power subject to a unity gain constraint for the target speaker in a reference microphone. Because the constraint set is limited to a single direction, steering vector errors (due to DOA mismatch) can lead to partial cancellation of the desired speech. MVDR beamformer is a more robust alternative to MPDR beamformer, which replaces the total power criterion with noise-plus-interference power. However, it requires an accurate estimate of the noise-and-interferer covariance matrix. Whereas, LCMV beamformer extends MVDR beamformer by introducing multiple linear constraints, such as, preserving a second speaker or imposing nulls toward dominant interfering source. While these additional constraints enable more control of interfering speakers, they reduce the degrees of freedom available for noise suppression, so the residual-noise performance

is often slightly worse than MVDR beamformer. Both MVDR and LCMV beamformers can be implemented either directly, by solving the constrained optimization with the current covariance and steering estimates, or by using a generalized sidelobe canceller architecture, which converts the constrained problem into an unconstrained one. Estimation of accurate steering vectors depends on modeling the full acoustic transfer functions (ATFs) between sources and microphones, which include delay, reverberation, microphone colouration, and body-related filtering (e.g., from a head or device). Estimating the full ATFs accurately is challenging, especially in noisy and reverberant conditions. Therefore, to handle this problem, adaptive beamforming algorithms often use relative transfer functions (RTFs), which is commonly estimated using either the covariance subtraction method or the covariance whitening method.

Many work have also investigated adaptive beamforming for target speaker extraction, both without DNNs [115], and in combination with DNN [41], [116]–[118]. In [115], a two-stage framework is proposed, speaker extraction and ASR. At first the complex Gaussian mixture model is extended to estimate the masks for each target speaker, interfering speaker and background noise. These masks are used to compute three spatial covariance matrices, for each target, interfering and background noise. Then an MVDR beamformer is utilized, where a weight vector is computed that minimizes the combined power of the interfering speaker and background noise covariances while ensuring that source arriving from the target direction get passed without distortion. In [115], both beamforming and mask estimation is performed in loop. In [116], a DNN is used to estimate the time-frequency mask for the target speaker, from which spatial covariance matrices are computed and provided as input to the MVDR beamformer. Further in [117], the classic covariance-inversion step of MVDR is replaced with RNN, the traditional MVDR beamformer is transformed into a fully differentiable by replacing its non-differentiable matrix inversion and eigenvalue decomposition steps with two RNNs that directly predict frame-wise beamforming weights. In [118], a TCN-based architecture is used to estimate spatially informed masks, followed by a neural beamforming layer that refines the extracted signal, while [41] combines three modules together, a DNN-based DOA estimator, an adaptive beamformer, and a complex RNN-based denoising module to perform the target speaker in real-time.

2.1.2 *Independent component analysis (ICA)*

ICA is another approach to perform the speech separation [119]–[121]. It models the mixture signal as a linear combination of underlying latent components [122]–[124], which assumes each component is statistically independent of other. ICA exploits the statistical independence of the sources and therefore requires no prior knowledge of the array geometry or the DOA. The mixture signal $\mathbf{y}(n)$ is model as:

$$\mathbf{y}(n) \;=\; \mathbf{A}\,\mathbf{x}(n), \tag{2.1}$$

where $\mathbf{A}$ is the unknown mixing matrix and $\mathbf{x}(n) = [x_1(n), x_2(n), \dots]^\top$ is the source vector, whose components are assumed to be mutually statistically independent.

Under this assumption, the joint probability density function (PDF) can be defined as:

$$p\big(x_1, x_2, x_3, \dots\big) = p_1(x_1)\ p_2(x_2)\ p_3(x_3)\ \dots \tag{2.2}$$

where $p_i(\cdot)$ denotes the marginal PDF of the i-th source. In a noisy mixture (when the mixture consists of overlapping speech as well as background noise), it is assumed that at least one source has a non-Gaussian distribution, and the mixing matrix $\mathbf{A}$ is invertible. These assumptions are essential because, according to the central limit theorem, a linear mixture of independent signals tends to be more Gaussian than the individual sources. ICA separates the sources by maximizing non-Gaussianity [124].

ICA is also utilized for target speaker extraction in [27], [125]. For instance, in [125], the target speaker is extracted by combining ICA along with beamforming, while in [27] voice activity information derived from video signal is utilized with ICA to perform the speaker extraction. While ICA provides a feasible solution for source separation, its performance is often limited by the permutation ambiguity, which arises due to the lack of modeling inter-frequency dependencies among components. To address this limitation, IVA was introduced, which incorporates inter-frequency dependencies within each source. Unlike ICA, which assumes complete independence across all components, IVA preserves the statistical structure of individual sources across frequency bins. In the following section, we provide a brief overview of IVA algorithm.

2.1.3 *Independent vector analysis (IVA)*

IVA extends ICA by incorporating statistical dependencies within each source across frequency bins while preserving the mutual independence across different sources [126]–[129]. Specifically, IVA models each frequency domain source as a multivariate vector. For IVA, the joint distribution of all sources is approximated using the product of the marginal distributions of each multivariate source vector, which enables IVA to maintain intra-source dependencies while enforcing inter-source independence. Let I be the number of sources and F be the number of frequency bins, then

$$\mathbf{x}_i \ = \ \big[x_i^{(1)},\, x_i^{(2)},\, \dots,\, x_i^{(F)}\big]^\top, \qquad i = 1, 2, \dots, I, \tag{2.3}$$

IVA assumes these I source vectors are mutually independent while allowing the F components within each vector to remain statistically dependent. This modeling can be expressed by approximating the true joint PDF $p(\mathbf{x}_1, \dots, \mathbf{x}_I)$ with a product of multivariate PDFs that preserve the intra-source structure:

$$p(\mathbf{x}_1, \dots, \mathbf{x}_I) \ \approx \ \prod_{i=1}^{I} u_i\big(\mathbf{x}_i\big), \tag{2.4}$$

where $u_i(\mathbf{x}_i)$ is the multivariate marginal PDF for the i-th source vector. The frequency-wise separating matrices $\mathbf{G}^{(k)}$ with $k = 1, 2, .., F$ are then obtained by

minimizing the Kullback−Leibler (KL) divergence between the true joint PDF and the model in (2.4):

$$J_{div} = \mathrm{KL}\Big(p(\mathbf{x}_1, \ldots, \mathbf{x}_I) \,\Big\|\, \prod_{i=1}^{I} u_i(\mathbf{x}_i)\Big) = \int p(\mathbf{x}_1, \ldots, \mathbf{x}_I) \log \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_I)}{\prod_{i=1}^{I} u_i(\mathbf{x}_i)} \, d\mathbf{x}_1 \cdots d\mathbf{x}_I. \tag{2.5}$$

2.1.4 *Independent vector extraction (IVE)*

While ICA and IVA aim to estimate all latent sources simultaneously under the assumption that they are mutually independent, IVE aims at only extracting the target speaker from the mixture [54], [130]–[133]. Similar to ICA and IVA, IVE models the mixture $\mathbf{y}(n)$ as a linear combination of latent sources:

$$\mathbf{y}(n) = \mathbf{A}\,\mathbf{x}(n), \tag{2.6}$$

where $\mathbf{A}$ is the unknown mixing matrix and $\mathbf{x}(n) = [x_1(n), x_2(n), \ldots, x_I(n)]^\top$ is the source vector consisting of I sources. IVE aims to extract only the target source, $x_j(n)$ by estimating a single demixing vector $\mathbf{w}_{\mathrm{IVE}}$, such that:

$$\hat{x}_j(n) \;=\; \mathbf{w}_{\mathrm{IVE}}^{\top}\,\mathbf{y}(n) \tag{2.7}$$

$\mathbf{w}_{\mathrm{IVE}}$ is chosen to maximize the statistical independence between the estimated target speaker $\hat{x}_j(n)$ and interfering sources, by minimizing the KL divergence between the joint distribution $p(\hat{x}_j, \mathbf{z}_{\mathrm{int}})$ and the product of the marginal approximations. $\mathbf{z}_{\mathrm{int}}$ denotes the interfering sources of $\mathbf{x}(n)$. IVE can also be advantageous when the number of sources is not known in the mixture, as it does not require full demixing of all components. In [133], the DOA of the target speaker is used as auxiliary information for extracting the target speaker, while in [54], a combination of IVE is used along with a supervised speaker identification system to improve target speech extraction performance. Whereas in [132], an extension of IVE is proposed to extract the moving target speaker.

Despite the significant progress made using statistical signal processing, these approaches face several limitations. These approaches often rely on strong assumptions about the number of sources, independence of speakers, or prior knowledge about the direction of the target speaker or array geometry. In contrast, DNN-based approaches have gained significant attention in speech processing by enabling data-driven feature extraction and robust modeling of speech characteristics. DNN-based approaches learn feature representations directly from data, making them highly adaptable to diverse acoustic conditions. The following sections explore how DNNs work and how they have been utilized for target speaker extraction.

2.2 Approaches based on deep neural networks

In this section, we first discuss the general working principles of a deep neural network (DNN) in Section 2.2.1, followed by an overview of DNN-based SC-TSE algorithms in Section 2.2.2. We then discuss computing speaker embeddings from auxiliary information in Section 2.2.2.1 and methods for conditioning the DNN on the target speaker in Section 2.2.2.2. We further discuss the operating domains: the time-frequency and time domains in Section 2.2.2.3, and loss functions in Section 2.2.2.4.

2.2.1 *Deep neural network (DNN)*

DNNs have gained significant attention in recent years due to their outstanding performance across various fields, including image processing, video analysis, machine translation, and speech processing. Their success is largely attributed to their ability to model complex, non-linear relationships between inputs and outputs [134], [135]. By learning hierarchical representations from data, DNNs mimic human cognition, identifying relevant patterns in training data and using them to make predictions on unseen test data. A typical DNN [135] consists of three types of layers: an input layer, one or more hidden layers, and an output layer. The hidden and output layers are composed of numerous interconnected neurons, which serve as the core computational units of the network. While the input and output layers depend on the specific task, the number of hidden layers and neurons per layer are hyperparameters chosen during network design.

For instance, in a fully connected DNN [14], [15], [136]–[138], each input from the input layer is connected to each neuron in the first hidden layer, and each neuron in one hidden layer is connected to each neuron in the next layer. A single neuron may have multiple input and output connections, forming many-to-many relationships. For instance, if the inputs to a neuron are $d_1, d_2, \ldots, d_H$, and $\mathbf{d}$ is a vector containing these inputs, then its output can be expressed as the weighted sum of these inputs, processed through a non-linear activation function; i.e.,

$$z_{w,b}(\mathbf{d}) = f\left(\sum_{i=1}^{H} w_i d_i + b\right), \tag{2.8}$$

where $f(\cdot)$ represents the activation function, d_i and w_i are the input and corresponding weight of the i-th connection, respectively, and b is the bias. The weights determine the relative importance of each input, with higher weights indicating greater importance. The bias serves as an additional parameter that ensures the network can fit complex patterns. Both weights and biases are trainable parameters adjusted during learning. In supervised learning, DNNs are trained by minimizing the difference between their predicted and desired outputs. This optimization is typically achieved through backpropagation and gradient descent, guided by a loss function. During training, the network iteratively updates weights and biases to reduce prediction error, gradually refining its performance.

The DNN-based approaches rapidly gained significant attention with the success of deep clustering [137] and permutation invariant training (PIT) [5] for single-channel, speaker-independent speech separation, which also significantly helped in advancing research on target speaker extraction. In the next section, we discuss how DNN-based approaches are used for target speaker extraction utilizing reference speech as auxiliary information.

2.2.2 *SC-TSE algorithms*

Fig. 2.1 depicts the overall framework of an SC-TSE algorithm utilizing reference speech of the target speaker. SC-TSE algorithms typically consist of two networks: the auxiliary network (referred to as the speaker embedder) and the main network (referred to as the speaker separator). The input to the speaker embedder network is the reference speech of the target speaker, $a_j(n)$. The input to the speaker separator network is the mixture of the target and interfering speakers and background noise, $y(n)$. The separator network is informed by the speaker embedding $\mathbf{e}_j$ computed from the reference speech of the target speaker. The speaker separator network computes a mask, a multiplicative function, or a multi-frame filter to perform speaker extraction, e.g., in the time-frequency domain [14], [15], [20], [21], [47] or in the time domain [16]–[19], [22], [23]. The estimated target speaker is compared with the ground-truth target speaker, using a suitable loss function $\mathcal{L}(\cdot, \cdot)$.

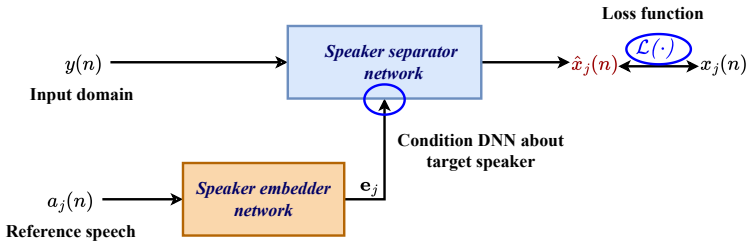

Fig. 2.1: Overall framework of DNN-based SC-TSE, consisting of a speaker separator network that is conditioned by a speaker embedding computed from reference speech of the target speaker.

When designing an SC-TSE algorithm, the following key components must be carefully considered:

- Speaker embedding representations: how to effectively obtain compact and discriminative features of the target speaker from the reference speech using the speaker embedder network (see Section 2.2.2.1).
- Integration of speaker embedding: how to utilize the obtained speaker embedding to condition the speaker separator network to effectively extract the target speaker from the mixture (see Section 2.2.2.2).
- Operating domains: the selection of appropriate input and output domains for the speaker embedder and separator networks, i.e., time-frequency domain and time domain (see Section 2.2.2.3).

- Loss function: how to compare the extracted target speaker with the ground-truth target speaker (see Section 2.2.2.4).
- Network architecture: how to choose efficient DNN architectures for the speaker embedder and speaker separator networks (see a separate Section 2.3).

2.2.2.1 *Speaker embedding representations*

The speaker embedding represents the unique characteristics of the target speaker obtained from the reference speech. The speaker embedder network computes a compact representation of the target speaker, which helps the speaker separator network to distinguish the target speaker from other sources, such as interfering speakers or background noise. Speaker embeddings have been extensively researched in speaker identification and verification tasks. Commonly used systems for computing speaker embeddings include the i-vector extractor [139], DNN-based speaker embedder networks, such as x-vector or d-vector [53], [140], and jointly learned speaker embedder networks [16], [17]. In this section, we briefly discuss these systems.

I-vector extractor

I-vectors [139] are fixed-dimensional feature vectors extracted from a given utterance using a Gaussian mixture model (GMM) whose parameters are constrained to a subspace (see Fig 2.2). This subspace is defined by the Universal Background Model (UBM), a GMM trained on a large dataset containing speech from many speakers, and a total variability subspace matrix, denoted by $\mathbf{T}$. In [139], the following model was proposed:

$$\mathbf{m} = \mu + \mathbf{T}\mathbf{s}, \tag{2.9}$$

where $\mathbf{m}$ and μ denote the mean super-vector of the utterance-specific GMM, and the mean super-vector of the UBM, respectively. $\mathbf{T}$ represents a low-rank rectangular matrix representing the bases spanning the subspace, while vector $\mathbf{s}$ is a latent variable drawn from a standard normal distribution. Both the UBM and $\mathbf{T}$ are pre-trained on a large dataset without using the speaker labels, where each utterance is considered as coming from a different speaker.

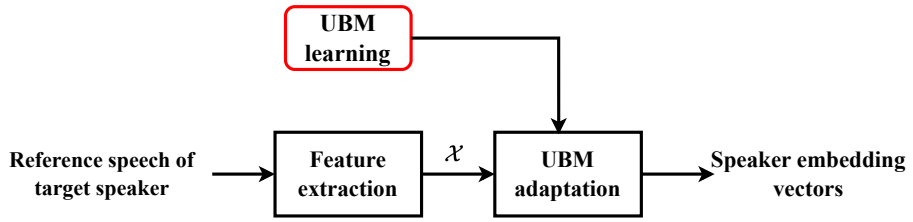

Fig. 2.2: Computation of target speaker embedding based on i-vector extractor. The red block is only considered during training.

Given a feature representation $\mathcal{X}$, such as Mel-Frequency Cepstral Coefficients (MFCCs) of the reference speech of the target speaker, the i-vector extractor computes the posterior distribution $p(\mathbf{s} \mid \mathcal{X})$, which models the uncertainty over the latent variable $\mathbf{s}$ conditioned on the observed feature under the total variability model. This model captures both speaker and session variability, factors that influence the acoustic features but are not related to the identity of the speaker, including microphone type, recording environment, or transmission channel. Due to the linear-Gaussian assumptions of the model, specifically, a Gaussian prior on $\mathbf{s}$ and a linear relationship between $\mathbf{s}$ and $\mathbf{m}$ (see (2.9)), the resulting posterior distribution $p(\mathbf{s} \mid \mathcal{X})$ is also Gaussian, which allows for efficient and closed-form computation of both the posterior mean and covariance. The i-vector is then computed as the posterior mean of $p(\mathbf{s} \mid \mathcal{X})$, corresponding to the maximum a posteriori (MAP) estimate. Due to their ability to capture both speaker and session variability, i-vectors have been widely used for speaker verification [141] and have also been utilized as speaker embeddings in SC-TSE algorithms [14], [142].

DNN-based speaker embeddings

Current state-of-the-art speaker identification and verification systems predominantly rely on DNN-based speaker embeddings. These embeddings can be either time-invariant [6] or time-varying [22], [143], representing the unique characteristics of a speaker. Among the most widely used speaker embedding extractors are x-vector [140] and d-vector [53]. DNN-based speaker embeddings are extensively employed in SC-TSE algorithms [15], [101], [102], [144], [145].

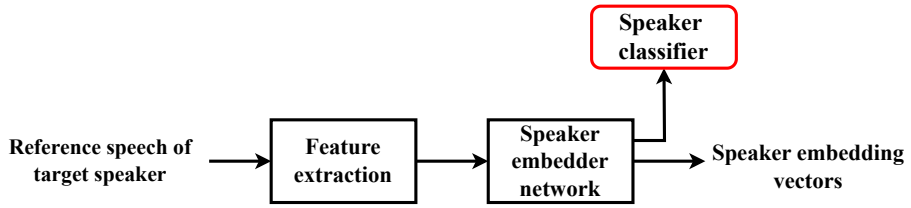

Fig. 2.3: Computation of DNN-based target speaker embedding. The red block is only considered during training.

The process of computing speaker embeddings using x-vector or d-vector extractors generally involves training a DNN-based speaker identification system. This system utilizes a pooling operation to transform a sequence of input features into a compact vector representation. The pooling operation can be performed using simple statistical methods, such as computing the mean and standard deviation [140], or by employing DNN-based architectures, such as long short-term memory (LSTM) layers [53] or attention mechanisms [146]. Several loss functions [53], [147]–[150] have been explored to train these systems. Commonly used loss functions include the cross-entropy loss [147] and its enhanced variants [148], [149], which focus on optimizing the inter-class separability among speaker representations. The triplet loss [150], which directly optimizes the relative distances between embeddings of the same and different speakers, and the generalized end-to-end (GE2E) loss [53],

which optimizes intra-speaker variability and maximizes inter-speaker separation without requiring explicit speaker labels per batch. In a speaker identification system, the resulting embeddings are passed to a classification layer to predict the speaker categories. However, for target speaker extraction, the final classification layers of the speaker identification system are removed, and only the speaker embedding obtained from the activation function of one of the last layers is used (see Fig. 2.3).

Jointly learned speaker embeddings

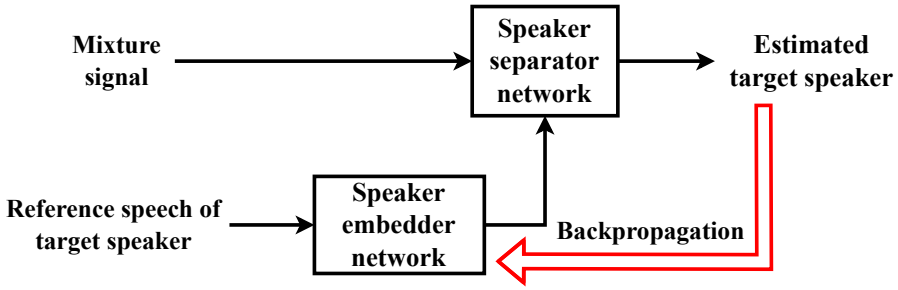

Fig. 2.4: Computation of target speaker embeddings by jointly training the speaker embedder and speaker separator networks. The red part is only considered during training.

Instead of using DNN-based speaker embeddings, trained for speaker identification tasks, in SC-TSE algorithms, an alternative approach is to train the speaker embedder network alongside the speaker separator network using a weighted combination of loss functions (see Fig. 2.4). this ensures that the target speaker embedding are specifically optimized for target speaker extraction. Jointly learned speaker embeddings are recently more often considered in SC-TSE [16]–[19], [21], [103].

2.2.2.2 *Methods for conditioning the speaker separator network*

The performance of an SC-TSE algorithm highly depends on how the target speaker embedding guides the separator network to discriminate between target speaker and interfering sources to extract the target speaker from the mixture. In the following, we will discuss several conditioning methods that have been explored for incorporating speaker embedding into the separator network.

Concatenation method

The concatenation method is one of the simplest approaches for incorporating target speaker information into the speaker separator network. In this method, the speaker embedding $\mathbf{e}_j \in \mathbb{R}^D$ is concatenated with the input to a specific layer within the

speaker separator network, referred to as the adaptation layer q. The processing of each layer $a \in \{1, 2, \ldots, A_l\}$ in the speaker separator network can be expressed as:

$$\mathbf{\Gamma}_a = \begin{cases} \sigma_a(L_a([\mathbf{\Gamma}_{a-1}, \mathbf{e}_j]; \theta_a)), & \text{for } a = q, \\ \sigma_a(L_a(\mathbf{\Gamma}_{a-1}; \theta_a)), & \text{for } a \neq q, \end{cases} \tag{2.10}$$

where $\mathbf{\Gamma}_{a-1}$ and $\mathbf{\Gamma}_a$ denote the input and output of the a-th layer, respectively. $L_a(\mathbf{\Gamma}_{a-1}; \theta_a)$ denotes the transformation applied by the a-th layer, parameterized by θ_a, where σ_a is the activation function. The concatenation method is a popular choice due to its simplicity and successful application in SC-TSE algorithms [15], [16], [105].

In the concatenation method, the selection of the adaptation layer is crucial. If the speaker embedding is concatenated too early, the speaker separator network may struggle to retain target speaker information in deeper layers, while if the concatenation occurs too late, the separator network may fail to effectively utilize the discriminative feature of the target speaker embedding.

Factorized layer method

The factorized layer-based method offers an alternative approach for incorporating the speaker embedding into the speaker separator network [151]. Unlike the concatenation method, which simply concatenates the speaker embedding with input features, this method modifies the speaker separator network architecture by replacing the selected layer with a set of multiple sub-layers. The output of the factorized layer is computed as a weighted combination of the outputs from all sub-layers. Following the previous notation, let F_{sub} denote the number of sub-layers and q be the factorized layer. The processing of each layer in the speaker separator network can be expressed as:

$$\mathbf{\Gamma}_a = \begin{cases} \sigma_a\left(\sum_{f=0}^{F_{\text{sub}}-1} \lambda_j^{(f)} L_a(\mathbf{\Gamma}_{a-1}; \theta_a^{(f)})\right), & \text{for } a = q, \\ \sigma_a\left(L_a(\mathbf{\Gamma}_{a-1}; \theta_a)\right), & \text{for } a \neq q, \end{cases} \tag{2.11}$$

where $\lambda_j^{(f)}$ and $\theta_a^{(f)}$ denote the scalar weight assigned to the f-th sub-layer and the parameters of the f-th sub-layer, respectively. The sub-layer weights $\lambda_j^{(f)}$ are computed from the speaker embedding via a learned linear transformation followed by a softmax operation, which can be expressed as:

$$\boldsymbol{\lambda}_j = \text{softmax}(\mathbf{W}_{\text{sl}} \mathbf{e}_j + \mathbf{b}_{\text{sl}}), \tag{2.12}$$

where $\mathbf{W}_{\text{sl}}$ and $\mathbf{b}_{\text{sl}}$ are the weight matrix and bias vector of the linear layer, respectively, and $\boldsymbol{\lambda}_j = [\lambda_j^{(0)}, \lambda_j^{(1)}, \ldots, \lambda_j^{(F_{\text{sub}}-1)}]^\top$ represents the contribution of each sub-layer.

While this method allows target speaker embeddings to have a strong influence on speaker extraction performance [152], it also provides flexibility in extracting different target speakers by adjusting the weights assigned to each sub-layer. However,

incorporating a large number of sub-layers within the factorized layer can lead to increased computational and memory costs, making it more resource-intensive.

Multiplication method

The multiplication method is another alternative introduced in [14], where the outputs of the selected layer in the speaker separator network are element-wise multiplied with the target speaker embedding. This method is simpler compared to the factorized layer method while still having a strong influence on the speaker separator network. However, to use this method, the dimensions of the target speaker embedding and the output of the selected layer need to be the same. Let q be the selected layer, following the previous notation, the processing of each layer in the speaker separator network can be expressed as:

$$\mathbf{\Gamma}_a = \begin{cases} \sigma_a(\mathbf{e}_j \odot L_a(\mathbf{\Gamma}_{a-1}; \theta_a)), & \text{for } a = q, \\ \sigma_a(L_a(\mathbf{\Gamma}_{a-1}; \theta_a)), & \text{for } a \neq q, \end{cases} \tag{2.13}$$

where $\odot$ denotes the point-wise multiplication operation. This method can be especially helpful in achieving low complexity compared to concatenation and factorized layer methods [153].

FiLM method

Based on the multiplication method, feature-wise linear modulation (FiLM) has been proposed in [154], which introduces a bias vector in addition to the multiplicative modulation. Let q be the selected layer, following the previous notation, the processing of each layer in the speaker separator network can be expressed as:

$$\mathbf{\Gamma}_a = \begin{cases} \sigma_a(\mathbf{e}_j^{(mul)} \odot L_a(\mathbf{\Gamma}_{a-1}; \theta_a) + \mathbf{e}_j^{(add)}), & \text{for } a = q, \\ \sigma_a(L_a(\mathbf{\Gamma}_{a-1}; \theta_a)), & \text{for } a \neq q, \end{cases} \tag{2.14}$$

where $\mathbf{e}_j^{(mul)}$ and $\mathbf{e}_j^{(add)}$ denote the multiplicative modulation vector and additive modulation vector, respectively. Compared to the multiplication method, the FiLM method offers more flexibility by utilizing both feature amplification and shifting.

Attention method

In all previously discussed methods, the same speaker embedding is used for all time frames of the mixture. However, the most relevant information from the reference speech may vary depending on the specific time frame of the mixture. For example, if the current time frame in the mixture contains vowel sounds, it could be more effective to extract embeddings from the reference speech where the target speaker is also pronouncing vowels. This adaptive selection of relevant information can be achieved through an attention mechanism, as proposed in [155]. Following the previous notation, the processing of each layer in the speaker separator network can be expressed as:

$$\mathbf{\Gamma}_a = \begin{cases} \sigma_a(\mathbf{p}_{attn} \odot L_a(\mathbf{\Gamma}_{a-1}; \theta_a)), & \text{for } a = q, \\ \sigma_a(L_a(\mathbf{\Gamma}_{a-1}; \theta_a)), & \text{for } a \neq q, \end{cases} \tag{2.15}$$

$$\mathbf{p}_{attn} = \mathbf{w}_{attn}^{\top} \mathbf{E}_j, \tag{2.16}$$

$$\mathbf{w}_{attn} = \text{softmax}\big(\mathbf{E}_j L_a(\mathbf{\Gamma}_{a-1}, \theta_a)^{\top}\big), \tag{2.17}$$

where $\mathbf{E}_j \in \mathbb{R}^{N_e \times D}$ denotes the dynamic speaker embedding with N_e the number of time frames in the reference speech and D the speaker embedding dimension. The attention weights $\mathbf{w}_{attn}$ are computed by measuring similarity between each frame in the reference speech and each frame in the mixture representation. The resulting attended embedding $\mathbf{p}_{attn}$ is obtained by weighting the dynamic embeddings according to these attention scores. Due to the influence of attention weights, the target speaker embeddings capture more information from reference speech segments that are more similar to the current mixture representation. Instead of incorporating attention-based dynamic speaker embeddings using multiplication as in (2.15), they can also be incorporated using concatenation, summation, or a factorized layer.

Throughout this thesis, we focus only on using the concatenation method to condition the speaker separator network.

2.2.2.3 *Operating domains for SC-TSE algorithms*

Target speaker extraction can be performed in either the time-frequency domain or the time domain. In the following sections, we will discuss both domains.

Target speaker extraction in the time-frequency domain
In the time-frequency domain, speaker separator networks commonly rely on the short-time Fourier transform (STFT) to represent the mixture signal [15], [102], [156]. Considering the scenario as in (1.1), in the STFT domain the mixture signal $Y(k,l)$, with k and l denoting the frequency index and the time frame index, respectively, is given by

$$Y(k,l) = \sum_{i=1}^{I} X_i(k,l) + V(k,l), \tag{2.18}$$

where $X_i(k,l)$ denotes the speech component corresponding to the i-th speaker and $V(k,l)$ denotes the noise component. The speaker separator network aims at estimating the speech component $\hat{X}_j(k,l)$ corresponding to the target speaker (see Fig. 2.5) by multiplying the mixture signal with a (real-valued or complex-valued) mask $M_j(k,l)$, i.e.,

$$\hat{X}_j(k,l) = M_j(k,l) Y(k,l). \tag{2.19}$$

When estimating a real-valued mask, the speaker separator network typically takes the magnitude of the STFT $|Y(k,l)|$, along with the target speaker embedding as input. This mask is applied to the magnitude of the mixture STFT to estimate the

magnitude of the target speaker. However, because only the magnitude is used, the network does not estimate the phase of the target speaker, which is crucial for an accurate time-domain reconstruction of the target speaker signal. As a result, the target speaker signal is reconstructed using the estimated magnitude combined with the mixture phase, which can introduce artifacts. Alternatively, when estimating a complex-valued mask, the speaker separator network typically takes both the magnitude and phase of the mixture STFT, along with the speaker embedding as input as in [157], [158]. This enables the simultaneous estimation of both magnitude and phase of the target speaker, leading to more accurate extraction and natural-sounding reconstructions of the target speaker signal compared to estimating the real-valued mask [158].

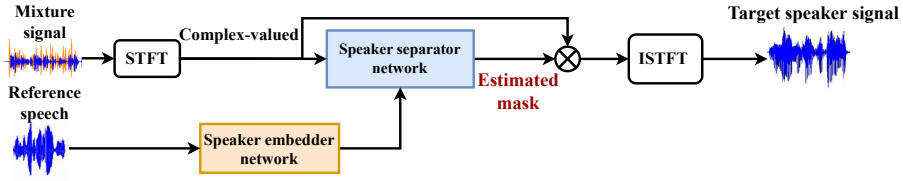

Fig. 2.5: Block diagram of target speaker extraction to estimate a (real-valued or complex-valued) mask in the time-frequency domain.

Target speaker extraction in the time domain

In the time domain, the speaker separator network takes a mixture signal along with the target speaker embedding as input [16]–[19], [22], [23]. Instead of using a predefined time-frequency transformation of the mixture signal, the separator network learns feature representations directly from the mixture signal. SC-TSE in the time domain has gained significant attention due to its ability to capture fine-grained temporal details while avoiding the challenges associated with phase estimation in the time-frequency domain (in case of real-valued mask estimation).

A time-domain-based speaker separator network (see Fig. 2.6) typically consists of a learnable speech encoder, which transforms the mixture signal into a latent representation (learned features or embedding coefficients), a mask estimator, which estimates a multiplicative function to extract the target speaker, and a speech decoder, which reconstructs the target speaker signal from the modified latent representation.

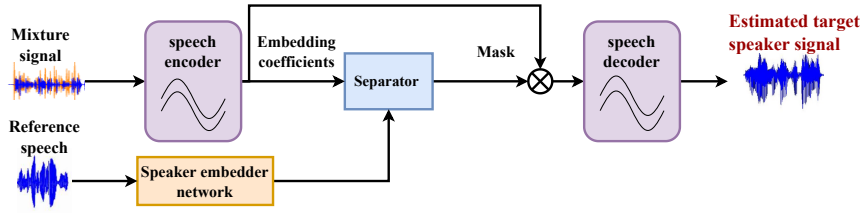

Fig. 2.6: Block diagram of target speaker extraction in the time domain, consisting of a speaker embedder, a speech encoder, a separator, and a speech decoder.

Target speaker extraction in the time domain has been shown to outperform the speaker extraction performance compared to the time-frequency domain-based speaker extraction [16], [18]. However, large amounts of training data are required for effective generalization. In addition, time-domain-based SC-TSE algorithms often require more complex architectures compared to time-frequency domain-based algorithms.

In Chapter 3 we will consider SC-TSE algorithms in the time-frequency domain using a real-valued mask, while in Chapter 4 we will consider SC-TSE algorithms in the time domain. In Chapter 5 we will subjectively evaluate the performance of a time-frequency domain-based SC-TSE algorithm and a time-domain-based SC-TSE algorithm.

2.2.2.4 *Loss functions*

To train the speaker separator network, either using a pre-trained speaker embedder network or jointly with the speaker embedder network, a loss function is required that measures the difference between the predicted target speaker signal and the ground-truth target speaker signal. It should be noted that loss functions are not limited to the domain in which the network operates. A loss function defined in the time domain can be utilized to train the network operating in the time-frequency domain. Conversely, time-domain network can be trained using time-frequency domain loss functions. In the following, we briefly review commonly used loss functions for target speaker extraction.

Loss functions in time-frequency domain
A commonly used loss function for SC-TSE algorithms in the time-frequency domain is the mean-square error (MSE) loss, defined as:

$$\mathcal{L}_{\text{MSE}} = \sum_{k,l} \Big(|\hat{X}_j(k,l)| - |X_j(k,l)| \Big)^2, \tag{2.20}$$

where $|X_j(k,l)|$ denotes the magnitude of the STFT of the ground-truth target speaker signal.
A more auditory-inspired loss function was proposed in [159] as a weighted sum of a magnitude spectrum loss and a complex spectrum loss. The magnitude spectrum loss focuses on the difference between the compressed magnitudes of the ground-truth and estimated target speaker signals, while the complex spectrum loss focuses on the difference between the compressed complex spectra. The loss function is defined as:

$$\mathcal{L}_{\text{AUD}} = \sum_{k,l} \Big(|\hat{X}_j(k,l)|^p - |X_j(k,l)|^p \Big)^2 \; + \; \lambda_{\text{AUD}} \sum_{k,l} \Big(\hat{X}_j(k,l)^p - X_j(k,l)^p \Big)^2, \tag{2.21}$$

where p denotes a compression factor ($0 < p < 1$), which applies a power-law compression to both magnitude and complex spectra and partially balances the im-

portance of quieter and louder sounds. λ_{AUD} denotes a weighting hyperparameter that balances the contribution of the complex spectrum loss relative to the magnitude spectrum loss. Both losses discussed here are computed before applying the inverse STFT (ISTFT). However, similar loss functions can also be applied after performing ISTFT, similarly as in [160].

Loss functions in time domain

The scale-invariant signal-to-distortion ratio (SI-SDR) [7], [161] is one of the most commonly used time-domain quality metrics for evaluating the performance of speech separation, enhancement and SC-TSE algorithms (see Section 2.4.1), its negative version is typically utilized as the loss function. SI-SDR represents the signal reconstruction error between the estimated target speaker signal and the scaled ground-truth target speaker signal, which is defined as:

$$\mathcal{L}_{SI\text{-}SDR}(x_j, \hat{x}_j) = -10 \log_{10} \frac{\sum_n \left|\alpha\, x_j(n)\right|^2}{\sum_n \left|\alpha\, x_j(n) - \hat{x}_j(n)\right|^2},$$

$$\alpha = \frac{\sum_n \hat{x}_j(n)\, x_j(n)}{\sum_n x_j^2(n)}. \tag{2.22}$$

Various other loss functions have been explored, mainly in the context of speech separation and speech enhancement. These include combinations of time-frequency domain and time-domain loss functions, as well as hybrid approaches that compute a loss function on intermediate feature representations, as proposed in [162]. Additionally, perceptual loss functions, such as those proposed in [163]–[165], have been investigated to improve the perceptual quality of enhanced speech. In [163], the conventional sample-wise MSE loss is replaced with a learned, differentiable representation of the PESQ metric [166] in a two-stage framework: first, a DNN is trained using MSE loss to predict true PESQ scores from paired ground-truth and enhanced spectrograms, then freeze this DNN and fine-tune a pre-trained enhancement algorithm by directly maximizing the predicted PESQ. In [164], a non-intrusive DNN is trained to predict the PESQ score, using its MSE against true PESQ as a reference-free perceptual loss and combining it with separate MSE terms on both speech and noise estimates, while recently, a differentiable PyTorch implementation of PESQ (referred to as torch-pesq) along with the SI-SDR loss (see (2.22)) is used train the DNN for speech enhancement algorithm in [165]. These alternative loss functions can also be considered for training SC-TSE algorithms.

2.3 Existing SC-TSE algorithms

In this section, we discuss many existing SC-TSE algorithms, some of them we will use as baseline systems to compare our proposed architectures.

As discussed in Section 1.2, SC-TSE algorithms can effectively leverage architectures originally developed for speech enhancement and speech separation. Early SC-TSE algorithms [14], [15] primarily utilized recurrent neural network (RNN)-based architectures, such as LSTMs [167] and GRUs [168], due to their ability to capture long-term dependencies in sequential data. However, with advancements in DNNs, more scalable alternatives emerged, such as TCNs [7], [16], [42], [144]. These architectures demonstrated comparable effectiveness to LSTMs while offering parallelization capabilities. Several other architectures have been also explored, such as dual-path RNNs [143], [169].

As described in Section 2.2.2, SC-TSE algorithms typically consist of two networks: a speaker embedder network and a speaker separator network. While the speaker separator network can directly adopt architectures designed for speech enhancement and speech separation, additional consideration needs to be taken when selecting an architecture for the speaker embedder network, particularly when using jointly learned speaker embeddings. A critical challenge is determining the optimal integration layer in the speaker separator network, incorporating the target speaker embeddings extracted from the reference speech. Early speaker embedder networks relied on relatively simple architectures, such as a few fully connected layers with a non-linearity or a single recurrent layer [53]. More recently, advanced architectures, such as Residual Neural Networks (ResNets) [17], [22], [156], TCNs [170], and attention-based network [171] have also been employed to compute the target speaker embeddings.

For the integration of target speaker embeddings into the speaker separator network, earlier layers are generally preferred, following any of the methods discussed in Section 2.2.2.2. In LSTM-based architectures, target speaker embeddings are typically incorporated at the second layer of the speaker separator network [14]. In TCN-based architectures, target speaker embeddings are usually incorporated after the first repetition of convolutional blocks [16]. Some SC-TSE approaches further enhance target speaker awareness by incorporating speaker embeddings at multiple layers of the separator network [172], incorporating multiple representations of speaker embeddings [173], and utilizing accumulated speaker embeddings [174].

SpeakerBeam

SpeakerBeam [14], [175] was initially introduced for multi-channel target speaker extraction, but its core concept is equally applicable to single-channel scenarios. SpeakerBeam was one of the first algorithms to propose DNN-based target speaker extraction conditioned on reference speech of the target speaker. This algorithm introduced three distinct methods for conditioning of target speaker information into the separator network: concatenation, factorized-layer, and multiplication. To compute target speaker embeddings, it explored both i-vectors as well as a sequence summarization network, with and without attention mechanism. For the speaker separator network, bidirectional LSTMs were employed to estimate a real-valued mask in the time-frequency domain. The training process utilized the MSE loss function (see (2.20)) along with a deep clustering regularization. Additionally, to compute the MSE, the time-frequency bins were weighted based on phase differences between the ground-truth target speaker and mixture signals. Furthermore,

SpeakerBeam explored a multi-task training objective by extending the separator network to include two output layers, one for predicting a mask and another for generating embeddings for deep clustering. To assess the effectiveness of SpeakerBeam for generalization to unseen speakers, the speaker extraction performance was compared against an approach where a separate DNN or a dedicated layer within a DNN was trained for each target speaker [63]. The results demonstrated the superior generalization capability of SpeakerBeam to open-set test conditions.

To further enhance target speaker extraction performance, jointly learned target speaker embeddings were introduced in [176]. This was later extended in [177] by integrating joint training with an ASR system. Finally, in [18] this algorithm was adapted to the time domain by using a TCN-based architecture instead of LSTM-based architecture for the speaker separator network.

VoiceFilter

VoiceFilter [15] was one of the first SC-TSE algorithms specifically designed for single-channel scenarios. The algorithm employed a pre-trained speaker identification system based on d-vectors to extract target speaker embeddings from the reference speech of the target speaker. The speaker embeddings were integrated into the speaker separator network using the concatenation method. The speaker separator network combined convolutional neural networks (CNNs) and LSTMs to estimate a real-valued mask in the time-frequency domain. The CNN layers were responsible for extracting local features from the magnitude spectrogram of the mixture, while the LSTM layers captured temporal dependencies. The training process utilized the power-law compressed reconstruction loss function (see (2.21)). Experimental results demonstrated that this algorithm outperformed the same architecture trained with a permutation-invariant training (PIT) loss for speech separation.

VoiceFilter was further extended to various applications, e.g., to improve streaming speech recognition [178], as a front-end for speaker verification [179], and to handle multi-speaker auxiliary information [180]. In Chapter 3 we will use the VoiceFilter as a baseline algorithm and improve its performance by introducing customization to the LSTM cells in the speaker separator network.

SpEx

SpEx [16] was one of the first SC-TSE algorithms to introduce a multi-scale feature representation for the mixture signal by utilizing a multi-scale encoder and to estimate the target speaker in the time domain. For the speaker separator network, the SpEx architecture utilized a multi-scale speech encoder, consisting of CNN layers to capture different time resolutions of the mixture signal, a TCN-based separator for estimating a multiplicative function to extract the target speaker, and a multi-scale speech decoder with deconvolutional layers to reconstruct the target speaker signal in the time domain. The speaker embedder network was based on a bidirectional LSTM and operated in the time-frequency domain. The speaker embedder and separator networks were jointly optimized for speaker extraction and identification tasks. SpEx was trained using a weighted combination of loss functions, i.e., multi-scale SI-SDR for the speaker separator network and cross-entropy loss (CE) for the speaker embedder network. Experimental results demonstrated that SpEx outper-

formed baseline time-frequency domain-based algorithms SpeakerBeam [175] as well as its variant proposed in [181], with the multi-scale speech encoder yielding better performance compared to the single-scale version. Furthermore, joint optimization of the speaker embedder and separator networks enhanced the overall performance, and a longer segment of reference speech led to improved speaker extraction results.

The SpEx algorithm was extended in [17] by jointly training the speaker embedder and separator networks in the time domain, ensuring that both components operated consistently within the same domain. The LSTM-based architecture for the speaker embedder network was replaced with a ResNet-based architecture, while the same TCN-based architecture was utilized for the speaker separator network. In addition, a weight sharing mechanism was adopted for the multi-scale speech encoder and the speaker embedder network. Experimental results demonstrated that the extended algorithm, referred to as SpEx+, improved the speaker extraction performance. In Chapter 4 we will use the SpEx+ as a baseline algorithm and improve its performance through a novel architecture that combines TCN with conformer blocks.

Deep extractor network

The Deep Extractor Network (DENet) [182] adopted a different approach for the target speaker extraction in the time-frequency domain to estimate a real-valued mask. First, the speaker embedder network computed a high-dimensional embedding for each time-frequency bin, encoding information about the target speaker in that bin. Second, the embedding space was modified by the second stage of the LSTM-based speaker separator network, which was conditioned on the speaker embedding extracted from the reference speech. This transformation restructured the embeddings such that time-frequency bins dominated by the target speaker were mapped to a distinct region within the embedding space. Third, the mask was estimated by measuring the similarity between the transformed embeddings and a canonical extractor embedding. The position of the canonical extractor was established based on the distribution of target speaker embeddings observed in the training data. During training, DENet employed an MSE loss function that emphasizes time-frequency regions dominated by the target speaker. Experimental results demonstrated that DENet outperformed the oracle Deep Attractor Network baseline proposed in [183] even when using very short reference speech segment.

Speaker inventory

Unlike conventional SC-TSE algorithms, which condition the speaker separator network with information solely about the target speaker, the speaker inventory algorithm [155] also incorporated information about potential interfering speakers present in the mixture, assuming reference speech of potential interfering speakers is available. An attention-based mechanism was employed to condition the separator network on both the target and interfering speakers, enhancing its ability to differentiate between them. Experimental results clearly showed the benefits of having access to reference speech of interfering speakers for target speaker extraction. However, such an approach is only useful in practice for scenarios where a predefined list of possible speakers is available, e.g., for meeting transcription.

ResNet-GRU
Inspired by VoiceFilter [15], a ResNet-GRU architecture for the speaker separator network was utilized instead of CNN-LSTMs for single-channel SC-TSE. The algorithm employed the same speaker embedder network and the same conditioning method as in [15] to extract the target speaker embedding. The separator network combined ResNet and GRUs to estimate a real-valued mask in the time-frequency domain. The training process utilized the SI-SDR loss function (see (2.22)). Experimental results demonstrated that this algorithm showed the similar performance as the VoiceFilter algorithm.

2.4 Performance measures

The performance of target speaker extraction can be assessed using three types of measures: objective measures (Section 2.4.1), subjective measures (Section 2.4.2), and downstream evaluation measures (Section 2.4.3).

2.4.1 *Objective measures*

Objective measures are widely used and frequently reported in the literature, primarily because they offer a straightforward and efficient means of evaluation. These measures can be classified into two categories (intrusive and non-intrusive), based on the availability of the ground-truth target speaker signal ($x_j(n)$) in the mixture. In this section, we discuss both intrusive and non-intrusive measures used to evaluate the performance of SC-TSE algorithms.

1. Intrusive measures, which require access to the ground-truth target speaker signal to assess the quality or intelligibility of the estimated target speaker signal, such as the signal-to-distortion ratio (SDR) [184] and its variants [161], short-time objective intelligibility (STOI) [185], and perceptual evaluation of speech quality (PESQ) [166].

   Signal-to-distortion ratio (SDR) and SI-SDR
   SDR [184] is one of the most commonly used evaluation metrics to assess the performance of SC-TSE algorithms. SDR is defined as:

$$\mathrm{SDR}(x_j, \hat{x}_j) = 10 \log_{10} \frac{\sum_n |h(n) \star x_j(n)|^2}{\sum_n |h(n) \star x_j(n) - \hat{x}_j(n)|^2},$$

$$\mathbf{h} = \arg\min_{\mathbf{h}} \sum_n |h(n) \star x_j(n) - \hat{x}_j(n)|^2, \tag{2.23}$$

   where $\mathbf{h}$ is a finite-length filter (up to a specified maximum length) [184] and $\star$ represents the convolution operator. SDR measures the distortion between the ground-truth target speaker signal and the estimated target speaker signal, while allowing the ground-truth to pass through a short linear filter to best align with the estimated signal. However, SDR can be too permissive, resulting

in a high score even when some frequency bands are completely missing as shown in [161]. A better alternative is SI-SDR [161], which permits only a scalar gain on the ground-truth signal, ensuring a uniform level adjustment before measuring distortion. Revisiting the definition of SI-SDR from (2.22):

$$\text{SI-SDR}(x_j, \hat{x}_j) = 10 \log_{10} \frac{\sum_n \left|\alpha\, x_j(n)\right|^2}{\sum_n \left|\alpha\, x_j(n) - \hat{x}_j(n)\right|^2},$$

$$\alpha = \arg\min_{\alpha} \sum_n \left|\alpha\, x_j(n) - \hat{x}_j(n)\right|^2, \tag{2.24}$$

where α can be obtained in closed form as:

$$\alpha = \frac{\sum_n \hat{x}_j(n)\, x_j(n)}{\sum_n x_j^2(n)}. \tag{2.25}$$

Short-time objective intelligibility (STOI)
STOI [185] is used to measure speech intelligibility. STOI operates by transforming both the ground-truth and estimated target speaker signals into time-aligned one-third octave band representations, denoted as $X_j^{(\text{oct})}(l, o)$ and $\hat{X}_j{}^{(\text{oct})}(l, o)$, respectively, where $o \in [0, 14]$ denotes the index of third-octave band. For each time frame and band index, a score is computed from a sliding window of 30 consecutive frames from both the ground-truth and the estimated target speech. The estimated window is first normalized to match the energy of the ground-truth window and then clipped to prevent large values. STOI is then computed as the average linear correlation coefficient between the reference and estimated window as:

$$\text{STOI}\big(x_j(n), \hat{x}_j(n)\big) = \frac{1}{LO} \sum_{l,o} \text{corr}\Big(X_j^{(\text{oct})}(l-29{:}l, o),\, \hat{X}_j'{}^{(\text{oct})}(l-29{:}l, o)\Big), \tag{2.26}$$

where $\hat{X}_j'{}^{(\text{oct})}(l-29{:}l, o)$ denotes the normalized and clipped estimated window, L denotes the total number of frames, and corr is the linear correlation coefficient.

Perceptual evaluation of speech quality (PESQ)
PESQ [166] is one of the widely used objective evaluation measures designed to approximate the subjective MOS for speech quality [186]. It was originally developed for narrow-band telephone speech (ITU-T Recommendation *P*.862) and later extended to wide-band services such as Voice over IP by *P*.862.2. The ground-truth and estimated target speaker signals are first time-aligned, then passed through a psychoacoustic auditory model, which transforms them into

internal representations of perceived loudness over time and frequency. These representations are further processed by a cognitive model, which estimates the perceptual impact of any distortions. The final output is a MOS-Listening Quality Objective (MOS-LQO) score, which typically ranges from 1 (very poor quality) to 4.5 (excellent quality), reflecting the perceptual similarity between the estimated and the ground-truth target speaker signals. In cases of extreme degradations, the score can drop slightly below 1.

2. Non-intrusive measures, which do not require the ground-truth target speaker signal, such as DNSMOS [187]. DNSMOS predicts the perceived quality of a speech signal using a DNN trained on a large dataset of speech samples rated by human listeners. It is mainly developed to predict subjective speech quality ratings, especially assessing the performance of the speech enhancement algorithms. It evaluates speech quality using three perceptual scores. First, the speech quality Mean Opinion Score (MOS), which evaluates the overall quality of the signal considering both distortions and artifacts, ranging from 1 (lowest quality) to 5 (excellent quality). Second, the background noise MOS, which measures the intrusiveness of the interfering sources for SC-TSE algorithms. Third, the overall MOS, a combined measure of the speech and noise MOS.

In this thesis, we will use SDR, SI-SDR, wide-band PESQ, and DNSMOS to evaluate the performance of SC-TSE algorithms in Chapter 3, while only SI-SDR is used in Chapter 4.

2.4.2 *Subjective measures*

Subjective measures are obtained through listening tests conducted with a group of listeners. In applications where the estimated target speaker is intended to be heard directly, such as in hearing aids, subjective performance measures provide the most reliable assessment of the performance of an algorithm.

To evaluate the speech quality and intelligibility of the estimated target speaker, several subjective evaluation measures can be used, including paired comparisons [188], and speech recognition threshold (SRT) [189]. In contrast, to assess the cognitive effort required by a listener, measures such as categorically scaled perceived listening effort [190] can be employed.

Paired comparisons are used to assess listener preferences between different versions of the same stimulus. In each trial, participants indicate which version made the target speaker easier to understand using a six-point rating scale. SRTs are used to measure speech intelligibility. The SRT is defined as the signal-to-noise ratio (SNR) at which a participant can correctly recognize 50% of the words uttered by the target speaker. In each trial, participants listen to mixtures of target and interfering speakers and processed mixture using the SC-TSE algorithms and select the recognized words from a matrix displayed on the screen. The SNR is adaptively adjusted based on their responses. Perceived listening effort is used to measure the subjective effort required to understand the target speaker. Participants rate this

effort on a 13-point categorical scale ranging from "no effort" (German: "mühelos") to "extreme effort" ("extrem anstrengend") (13 ESCU).

In this thesis, we will use all three subjective evaluation measures to evaluate the performance of SC-TSE algorithms in Chapter 5, where each of these measures is discussed in detail.

2.4.3 *Downstream evaluation measures*

When an SC-TSE algorithm is used as a preprocessing step for another task, e.g., ASR, speaker verification, the most effective way to evaluate its performance is by assessing the performance of that task. Downstream evaluations provide a more application-specific assessment of how well target speaker extraction improves the overall performance of the system, as in [191]. However, this type of evaluation comes with certain drawbacks, i.e., the evaluation may highly depend on the specific downstream system. If the downstream system undergoes modifications, the evaluation needs to be repeated, making it less flexible and more resource-intensive.

2.5 Summary

This chapter provided a comprehensive review of target speaker extraction approaches, covering both classical and DNN-based approaches. Early approaches such as beamforming, ICA, IVA, and IVE were discussed, highlighting their limitations in handling complex real-world scenarios. The transition to DNN-based approaches has significantly improved speaker extraction by using powerful feature learning capabilities. We provided a brief overview of methods for speaker embedding computation and integration to condition the DNN about the target speaker. We also discussed some commonly used loss functions along with operating domains for SC-TSE algorithms. Several baseline SC-TSE algorithms, such as SpeakerBeam, VoiceFilter, SpEx, and Deep Extractor, were reviewed, highlighting their key components. Finally, we also discussed some evaluation measures, including objective and subjective, typically used to evaluate the performance of the SC-TSE algorithms.

3

CUSTOMIZED LSTM CELLS FOR SPEAKER-CONDITIONED TARGET SPEAKER EXTRACTION IN TIME-FREQUENCY DOMAIN

As mentioned in Chapter 2, this chapter focuses on single-channel target speaker extraction in the time-frequency domain, where the speaker embedder and speaker separator networks are trained separately. Inspired by the working principle of LSTM cells [192], [193] and leveraging their success in many applications [100], [167], [194], this chapter proposes three novel variants of LSTM cells for the speaker separator network, each estimating a real-valued mask for target speaker extraction. The first proposed variant customizes only the forget gate of the LSTM cell, aiming at selectively retaining the target speaker information in the cell state. The second proposed variant extends the first variant by also customizing the input gate, aiming at more effective resetting of the cell state, allowing it to update with relevant information about the target speaker and simultaneously disregarding information from other sources. The third variant is inspired by [100], which introduces a novel auxiliary-modulation gate within the LSTM cell. While the information processing through the forget, input, and output gates remains unchanged initially, it is modulated during the resetting of the cell state. The purpose of the auxiliary-modulation gate is to enhance the learning of both long-term and short-term discriminative features of the target speaker with the help of the corresponding speaker embedding.

This chapter evaluates the effectiveness of the proposed variants of LSTM cells against standard LSTM cells, utilizing the VoiceFilter system as a baseline [15] (see Section 2.3). Additionally, their performance is compared against two-step methods based on BSS (see Section 1.2.1).

This chapter is partly based on the following publications:

[101] R. Sinha, M. Tammen, C. Rollwage, and S. Doclo, "Speaker-conditioned target speaker extraction based on customized LSTM cells," in *Proc. ITG Conference on Speech Communication*, VDE, Kiel, Germany, Sept-Oct. 2021, pp. 1–5

[102] R. Sinha, C. Rollwage, and S. Doclo, "Variants of LSTM cells for single-channel speaker-conditioned target speaker extraction," *EURASIP J. on Audio, Speech, and Music Processing*, vol. 2024, no. 1, pp. 1–13, 2024

The remainder of this chapter is organized as follows: Section 3.1 provides a general overview of SC-TSE algorithms in the time-frequency domain (briefly revisiting Section 2.2.2.3). Section 3.2 discusses the speaker embedder network used in this chapter and each proposed LSTM cell variant, after briefly reviewing the standard LSTM cells. Section 3.3 discusses the experimental setup, including the datasets, network architectures, training and evaluation hyperparameters, and the considered two-step baseline systems. Section 3.4 presents the experimental results and performance analysis of the considered SC-TSE systems. Finally, Section 3.5 provides a summary of this chapter.

3.1 SC-TSE in time-frequency domain

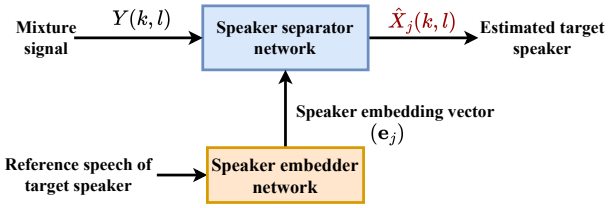

Fig. 3.1: Block-diagram of SC-TSE system operating in the time-frequency domain.

In this section, we briefly revisit target speaker extraction in the time-frequency domain from Section 2.2.2.3. Similar to the SC-TSE systems in [15], [156], we consider a speaker separator network that aims to extract the speech component $X_j(k,l)$ corresponding to the target (j-th) speaker by multiplying the microphone signal with a real-valued soft mask, i.e.,

$$\hat{X}_j(k,l) = M_j(k,l)Y(k,l). \tag{3.1}$$

The goal of the speaker separator network is to estimate the mask $M_j(k,l)$ using the target speaker embedding $\mathbf{e}_j$ (see Fig. 3.1) generated from the reference speech of the target speaker using the speaker embedder network. Finally, the time-domain signal is reconstructed by applying the inverse STFT to $\hat{X}_j(k,l)$ using a weighted overlap-add procedure.

3.2 Overview of SC-TSE system architecture

Similarly to [15], we use a pre-trained LSTM-based speaker identification system [53] as the speaker embedder network to generate the target speaker embedding from the reference speech of the target speaker. As discussed in Section 2.2.2.1, the speaker embedding guides the speaker separator network to estimate the target speaker from the mixture (see Fig. 3.2). We provide details of the speaker embedder

network in 3.2.1, and first provide a brief overview of standard LSTM cells before discussing the proposed LSTM cell variants in 3.2.2.

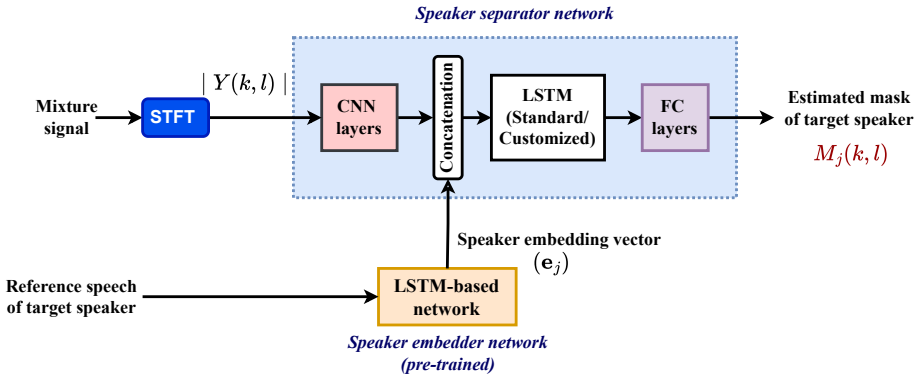

Fig. 3.2: An SC-TSE system utilizing standard or proposed customized LSTM cells for target speaker extraction in the time-frequency domain.

3.2.1 *Speaker embedder network*

In this chapter, we have used the same LSTM-based system [53] as the speaker embedder network (see Fig. 3.3). This system was originally trained for the speaker identification task, where variable-length speech utterances are mapped into fixed-dimensional vector representations (d-vectors) which capture the speaker identity. During training of this system for speaker identification [53], batches consisting of multiple speakers, each with several utterances, are processed. First, a d-vector for each utterance is extracted. Then, for each speaker, a centroid is computed by averaging their d-vectors within the batch, representing the speaker in the embedding space. The Generalized End-to-End (GE2E) loss function is then applied by constructing a similarity matrix using the cosine similarity between every d-vector and all speaker centroids, which encourages embeddings to cluster closely with their respective speaker centroids while distancing themselves from the centroids of other speakers.

During inference, when using this system as a speaker embedder network, we provide the reference speech of the target speaker as input and generate a fixed-dimensional embedding vector (see Fig. 3.3), which is used to guide the speaker separator network.

3.2.2 *Speaker separator network*

The speaker separator network aims at estimating a real-valued mask corresponding to the target speaker utilizing the target speaker embedding generated from the speaker embedder network. The speaker separator network in [15] uses a CNN-

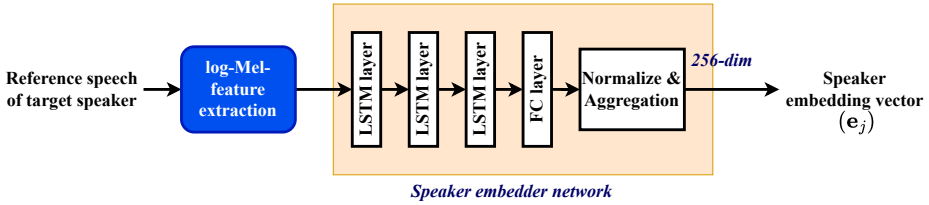

Fig. 3.3: Utilized pre-trained LSTM-based speaker embedder network.

LSTM architecture, where first convolutional operations are performed on the STFT magnitude of the microphone signal to obtain a transformed representation $\mathbf{r}$. The target speaker embedding $\mathbf{e}_j$ is then repeated and concatenated with this transformed representation along the frequency dimension and provided as input to the LSTM block for estimating the mask, i.e.,

$$M_j(k,l) = \boldsymbol{\phi}(\mathbf{r}, \mathbf{e}_j), \tag{3.2}$$

$$\mathbf{r} = \mathbf{g}(|Y(k,l)|), \tag{3.3}$$

where $\mathbf{g}(\circ)$ denotes the convolutional operations and $\boldsymbol{\phi}(\circ)$ denotes the LSTM block consisting of LSTM cells and fully connected layers. Instead of using standard LSTM cells, we propose three variants of LSTM cells that are customized for SC-TSE. Before discussing these variants, the working principle of the standard LSTM cell is briefly reviewed.

3.2.2.1 *Standard LSTM cells*

Fig. 3.4 depicts a standard LSTM cell [192], [193]. The working principle of an LSTM cell depends on its cell state and three gates: the forget gate, the input gate, and the output gate. The cell state behaves like the memory of the network, having a recursive property with the ability to retain information through time, while the different gates can add or remove information at each step t. In the following, the weight matrices and the bias vectors of the forget gate, the input gate, the output gate and the control update are denoted by $\mathbf{W}_f$, $\mathbf{W}_i$, $\mathbf{W}_o$, $\mathbf{W}_c$, and $\mathbf{b}_f$, $\mathbf{b}_i$, $\mathbf{b}_o$, $\mathbf{b}_c$, respectively. The current and previous cell states are denoted by $\mathbf{c}_t$ and $\mathbf{c}_{t-1}$, while the current and previous hidden states are denoted by $\mathbf{h}_t$ and $\mathbf{h}_{t-1}$.

As can be seen from (3.4) and Fig. 3.4, the input to each gate of the LSTM cell is the concatenation of the transformed representation $\mathbf{r}$ obtained from the CNN layers and the target speaker embedding $\mathbf{e}_j$. The recursive property allows the LSTM cell to store information from the previous state. The forget gate of the LSTM cell decides which information should be retained or disregarded based on the previous hidden state and the current input. The information is retained in the cell state if

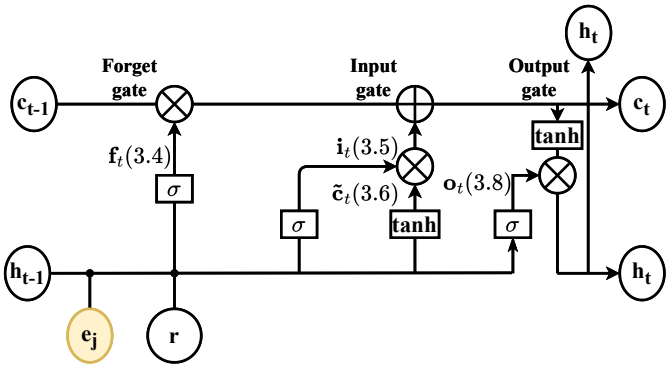

Fig. 3.4: Standard LSTM cell: the input to the forget gate $\mathbf{f}_t$ (3.4), input gate $\mathbf{i}_t$ (3.5) and output gate $\mathbf{o}_t$ (3.8) is the concatenation of the transformed representation $\mathbf{r}$ obtained from the CNN layers and the target speaker embedding $\mathbf{e}_j$. σ denotes the sigmoid activation function.

the output of the forget gate is close to 1, otherwise it is disregarded. The output of the forget gate is obtained as

$$\mathbf{f}_t = \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, (\mathbf{r}, \mathbf{e}_j)] + \mathbf{b}_f), \tag{3.4}$$

where σ denotes the sigmoid activation function.

The input gate has the ability to add new information to the cell state, but can not remove any information from it. It decides which information is getting updated and stored in the cell state, and its output is obtained as

$$\mathbf{i}_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, (\mathbf{r}, \mathbf{e}_j)] + \mathbf{b}_i). \tag{3.5}$$

The cell state behaves like the memory of the network, and is updated as

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c[\mathbf{h}_{t-1}, (\mathbf{r}, \mathbf{e}_j)] + \mathbf{b}_c), \tag{3.6}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t, \tag{3.7}$$

where $\odot$ denotes point-wise multiplication.

Finally, the output gate decides which part of the cell state is transferred to the next hidden state, i.e., the output of the output gate is obtained as

$$\mathbf{o}_t = \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, (\mathbf{r}, \mathbf{e}_j)] + \mathbf{b}_o). \tag{3.8}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t). \tag{3.9}$$

3.2.2.2 *Customized LSTM cells (F)*

As already mentioned, the forget gate is used to retain the relevant information and disregard irrelevant information, based on the previous hidden state and the current input. With the specific goal of speaker extraction in mind, intuitively the LSTM cell is supposed to learn to retain information related to the target speaker, while disregarding information unrelated to the target speaker, i.e., originating from the other speakers and background noise present in the mixture. However, since in practice this will not be perfectly achieved, we propose to customize the LSTM cell in order to only retain the target speaker information by changing the information processing through the forget gate. Fig. 3.5 depicts the proposed customized LSTM cell (F). Instead of considering the concatenation of the target speaker embedding $\mathbf{e}_j$ and the transformed representation obtained from the CNN layers $\mathbf{r}$, we only consider the target speaker embedding, i.e.,

$$\mathbf{f}_t = \sigma(\mathbf{W}_{ef}[\mathbf{h}_{t-1}, \mathbf{e}_j] + \mathbf{b}_{ef}), \tag{3.10}$$

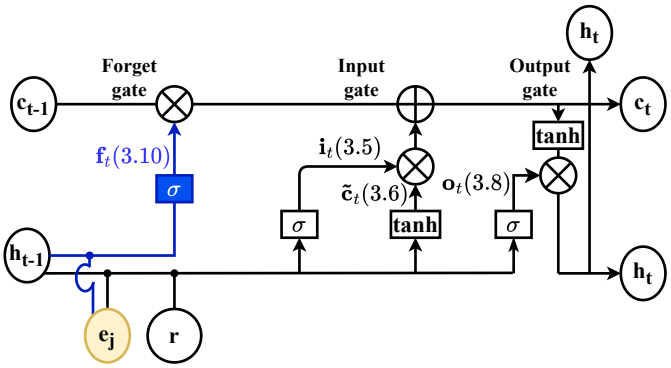

Fig. 3.5: Proposed customized LSTM cell (F): the input to the forget gate $\mathbf{f}_t$ (3.10) is only the target speaker embedding $\mathbf{e}_j$. σ denotes the sigmoid activation function.

where $\mathbf{W}_{ef}$ and $\mathbf{b}_{ef}$ denote the weight matrix and the bias of the customized forget gate. It should be noted that all other gates, i.e., the input and output gates, and the cell update remain the same as described in the previous Section 3.2.2.1. The motivation behind this proposed customization is to enable the forget gate to retain the information related to only the target speaker. The forget gate in (3.10) aims at mapping the target speaker close to 1. This allows the current cell state in (3.7) to retain the target speaker information by multiplying the previous cell state with a value close to 1, while disregarding the information related to the other speakers and background noise from the previous cell state $\mathbf{c}_{t-1}$, which directly affects the current hidden state and is utilized for computing the next hidden state.

3.2.2.3 *Customized LSTM cells (F + I)*

Based on our previous customization of only the forget gate, we also propose to customize the input gate along with the forget gate to simultaneously update and retain only the target speaker information. Fig. 3.6 depicts the proposed customized LSTM cell (F+I). Instead of using the concatenation of the target speaker embedding $\mathbf{e}_j$ and the transformed representation obtained from the CNN layers $\mathbf{r}$ as input to the forget gate and input gate, we only use the target speaker embedding $\mathbf{e}_j$ as input, i.e.,

$$\mathbf{i}_t = \sigma(\mathbf{W}_{ei}[\mathbf{h}_{t-1}, \mathbf{e}_j] + \mathbf{b}_{ei}), \tag{3.11}$$

where $\mathbf{W}_{ei}$ and $\mathbf{b}_{ei}$ denote the weight matrix and the bias of the customized input gate. As in the standard LSTM cell, the input to the output gate is the concatenation of the transformed representation $\mathbf{r}$ and the target speaker embedding $\mathbf{e}_j$. The motivation behind the proposed customization is to enable the forget gate to retain the information related to only the target speaker, while at the same time the input gate adds and updates the same information to the cell state. Hence, it is expected that information unrelated to the target speaker (other speakers, background noise) will be disregarded from the current cell state, which affects the current hidden state and is utilized for computing the next hidden state.

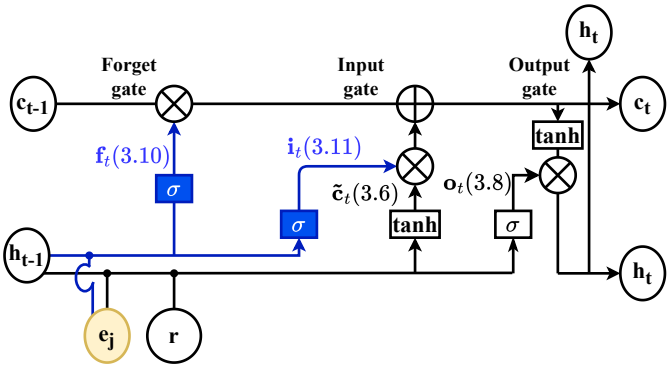

Fig. 3.6: Proposed customized LSTM cell (F+I): the input to the forget gate $\mathbf{f}_t$ (3.10) and the input gate $\mathbf{i}_t$ (3.11) is only the target speaker embedding $\mathbf{e}_j$. σ denotes the sigmoid activation function.

3.2.2.4 *Customized auxiliary-gated LSTM cells*

Inspired by the modified LSTM cell architecture presented in [100], in this customization we introduce a new gate within the LSTM cell, referred to as auxiliary-modulation gate. Fig. 3.7 depicts the proposed customized auxiliary-gated LSTM cell.

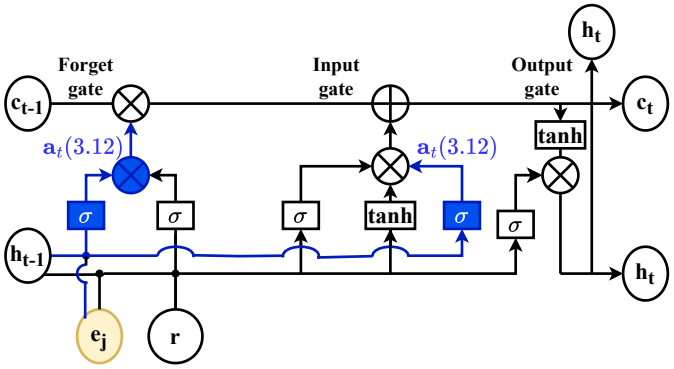

Fig. 3.7: Proposed customized auxiliary-gated LSTM cell: the input denotes the concatenation of the transformed representation $\mathbf{r}$ obtained from the CNN layers and the target speaker embedding $\mathbf{e}_j$, while the input to the auxiliary modulation gate $\mathbf{a}_t$ (3.12) is only the target speaker embedding $\mathbf{e}_j$. σ denotes the sigmoid activation function.

The auxiliary-modulation gate is defined as

$$\mathbf{a}_t = \sigma(\mathbf{W}_a[\mathbf{h}_{t-1}, \mathbf{e}_j] + \mathbf{b}_a), \tag{3.12}$$

where $\mathbf{W}_a$ and $\mathbf{b}_a$ denote the weight matrix and the bias of the auxiliary-modulation gate. It should be noted that the input to the auxiliary-modulation gate is only the target speaker embedding $\mathbf{e}_j$, while the input to all other gates (forget, input, output) is the concatenation of the transformed representation $\mathbf{r}$ and the target speaker embedding $\mathbf{e}_j$. The proposed auxiliary-modulation gate is part of both the forget gate and the input gate, hence leading to a direct impact on the cell state. The motivation behind this architecture is to modulate the information retained by the forget gate and the input gate by exploiting long-term and short-term discriminative features of the target speaker. After the modulation of information in the forget gate and the input gate, the cell state is updated as

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{a}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{a}_t \odot \tilde{\mathbf{c}}_t. \tag{3.13}$$

The motivation behind the proposed customization is to retain only the information related to the target speaker in the cell state at each step, while simultaneously disregarding the information unrelated to the target speaker.

3.3 Experimental setup

In this section, we present the experimental setup used for all considered SC-TSE systems (which we also referred to as one-step methods) and two-step target speaker extraction systems based on BSS. In Section 3.3.1, we discuss the speech datasets used for training and testing. In Section 3.3.2 and 3.3.3, we discuss the network archi-

tecture and the hyperparameters for the speaker embedder and separator networks of the SC-TSE systems. Finally, in Section 3.3.4 we discuss the network architecture and the hyperparameters for two two-step methods based on BSS.

3.3.1 *Speech datasets*

To generate the training, validation, and test data, we have used two different speech datasets with a sampling rate of 16 kHz, namely the Voxceleb dataset [195] and the Librispeech dataset [196]. We have used the Voxceleb dataset only to train the speaker embedder network, while we have used the Librispeech dataset employed by several existing SC-TSE systems [15], [101], [156], [197] to train, validate and test the speaker separator network.

Voxceleb dataset [195]: this dataset consists of more than 1 million utterances from more than 7000 speakers collected from YouTube. We have used the official training and validation split of the Voxceleb dataset only for training the speaker embedder network.

Librispeech dataset [196]: this dataset consists of 1000 hours of English speech with official training, validation, and test splits. The training set of the dataset is partitioned into 3 subsets, with approximate size 100 hours, 360 hours and 500 hours, respectively. In this chapter, we have only considered the 100 hours subset of the training set having 251 speakers to create the training data, while the official validation and test sets, each having 40 speakers have been used to create the validation and test data, respectively. To generate 2-speaker mixtures, we have used the same procedure as in [15], namely, randomly choosing two utterances from two different speakers and mix them together at 0 dB SNR. One speaker is considered as the target speaker, while the other speaker as the interfering speaker. A different utterance from the target speaker is used as the reference speech to obtain the target speaker embedding. A similar procedure has been followed to generate 3-speaker mixtures, where three different speakers are randomly chosen. After mixing both interfering speakers with the same power, the resulting mixture of interfering speakers has been mixed with the target speaker at 0 dB. All together, we have generated 160 hours of training data, 40 hours of validation data and 14 hours of test data for both 2-speaker and 3-speaker mixtures. We have used these training and validation data to train the speaker separator networks for the one-step methods and the BSS networks for the two-step methods, while the test data was used to evaluate all considered target speaker extraction systems.

3.3.2 *Speaker embedder network*

We have used the same speaker embedder network, i.e., the LSTM-based system, discussed in Section 3.2.1 for all considered SC-TSE systems. The speaker embedder network consists of 3 LSTM layers, each having 768 nodes. As input features the network uses 40-dimensional log-Mel-features, which are computed using an FFT size of 512, a Hann window with a frame length of 400 samples, and a frame shift of 160 samples. Given the log-Mel-features of the reference speech of the target speaker,

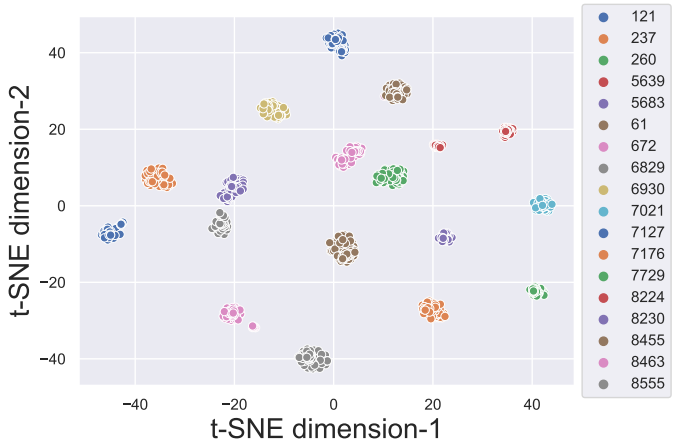

Fig. 3.8: Target speaker embeddings obtained from the reference speech of the target speaker on the Librispeech test set utilizing the speaker embedder network. 18 speakers are randomly selected for this visualization. t-SNE is used to reduce the embeddings in two dimensions. Each point on the plot represents one utterance of the target speaker, where colors represent the corresponding speakers.

the speaker embedder network generates a fixed 256-dimensional embedding vector. We have retrained the speaker embedder network on the Voxceleb dataset using stochastic gradient descent and a loss function referred to as generalized end-to-end (GE2E) loss presented in [53]. We have used the same parameters as in [53], namely an initial learning rate of 0.01, a batch size of 64, and a maximum of 1000 epochs. The visualization of the target speaker embeddings obtained from the reference speech of the target speaker using this pre-trained speaker embedder network is shown using t-SNE in Fig. 3.8.

3.3.3 *Speaker separator network*

Similar to the baseline system [15], the separator network of the SC-TSE systems consists of eight 2D dilated CNN layers, an LSTM layer and two fully connected (FC) layers. Each CNN layer is followed by a batch-normalization layer and a ReLU activation function. The parameters of the layers are shown in Table 3.1. The STFT magnitude of the mixture signal is computed using an FFT size of 512, a square-root Hann window with a frame length of 512 samples, and a frame shift of 256 samples. In total, we have trained 8 different speaker separator networks, depending on the LSTM cells used in the LSTM layer and whether unidirectional or bidirectional mode is considered:

- **Standard LSTM/BLSTM**: retrained baseline system [15] using standard LSTM cells.
- **Customized LSTM/BLSTM (F)**: proposed system presented using customized LSTM cells (F) described in Section 3.2.2.2.

Layer	Kernel size	Dilation	Filters/Nodes
Conv1	(1×7)	(1×1)	64
Conv2	(7×1)	(1×1)	64
Conv3	(5×5)	(1×1)	64
Conv4	(5×5)	(2×1)	64
Conv5	(5×5)	(4×1)	64
Conv6	(5×5)	(8×1)	64
Conv7	(5×5)	(16×1)	64
Conv8	(1×1)	(1×1)	8
LSTM	-	-	600
FC 1	-	-	514
FC 2	-	-	257

Table 3.1: Parameters of the separator network, consisting of eight CNN layers, an LSTM layer (standard or customized LSTM cells) and two fully connected layers.

- **Customized LSTM/BLSTM (F+I)**: proposed system using customized LSTM cells (F+I) described in Section 3.2.2.3.
- **Customized auxiliary-gated LSTM/BLSTM**: proposed system using customized auxiliary-gated LSTM cells described in Section 3.2.2.4.

All speaker separator networks have been trained using the SI-SDR loss function as discussed in Section 2.2.2.4 and the Adam optimizer [198] with a learning rate of 0.0002. We have used a batch size of 16 and fixed the maximum number of epochs to 50, while clipping the gradient norm to 10 and using an early stopping criterion of 7 epochs. The duration of the mixture signals was set to 4 seconds during training. To ensure the convergence for each system, training continued until the system achieved optimal generalization without overfitting or underfitting, monitored through both training and validation losses and controlled by early stopping.

3.3.4 *Two-step methods*

As additional baseline systems, we have also considered two-step target speaker extraction methods based on BSS (see Fig. 3.9). In the first step, the BSS network aims at estimating all individual speakers in the mixture, while in the second step, one of the output signals is selected as the target speaker based on the target speaker embedding. As the BSS network, we have considered the LSTM-based system in [5] and a system based on Conv-TasNet [7]. For the LSTM-based network, we have used a similar architecture as for the speaker separator network described in Section 3.3.3 but with three LSTM layers, each having 600 nodes. For the Conv-TasNet network architecture, we have used the same parameters as in [7] with causal mode using cumulative layer normalization. Both BSS networks have been trained using

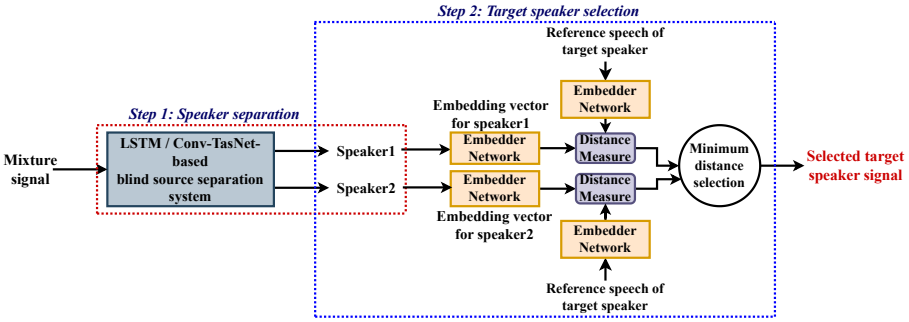

Fig. 3.9: Two-step method for target speaker extraction using BSS as the first step and target speaker selection as the second step. The embedder network used in the second step is the same speaker embedder network as in one-step SC-TSE.

the SI-SDR loss function with the same learning rate and batch size as for the speaker separator networks in Section 3.3.3. In order to select the estimated target speaker from the output of the BSS network, embedding vectors are computed for each estimated speaker and the reference speech of the target speaker using the speaker embedder network described in Section 3.3.2. The estimated target speaker is selected as the estimated speaker with the smallest $l2$-norm distance to the target speaker embedding. We have only trained and tested two-step target speaker extraction methods for 2-speaker mixtures (see results in Section 3.4.1).

3.4 Results and discussion

In this section, we compare the performance of the proposed SC-TSE systems using customized LSTM cells with the baseline system using standard LSTM cells [15] and two baseline two-step methods [5], [7]. As performance measures we have utilized the SDR, SI-SDR, wide-band PESQ (WB-PESQ), and DNSMOS (as discussed in Section 2.4.1). Additionally, the total number of parameters (computed using the torchinfo library of PyTorch) is also computed for all considered systems.

The results are reported in three Sections, depending on the data used for training and testing. In Section 3.4.1, both training and testing are performed on 2-speaker mixtures. To investigate the robustness against having more interfering speakers in the mixture during testing than during training, in Section 3.4.2, systems trained on 2-speaker mixtures are evaluated on 3-speaker mixtures. Finally, in Section 3.4.3 systems are evaluated on mixtures containing one, two or three speakers with and without background noise when performing training on multi-condition mixtures. In Section 3.4.1 and Section 3.4.2, the results are reported only in terms of SDR, while in Section 3.4.3 the results are reported in terms of all considered performance measures.

3.4.1 *Evaluation on 2-speaker mixtures*

For all considered systems, i.e., one-step methods using standard and customized LSTM cells (unidirectional and bidirectional mode), and baseline two-step methods (unidirectional mode), Table 3.2 shows the mean SDR and the total number of parameters for 2-speaker mixtures, where all systems were trained on 2-speaker mixtures. For the two-step methods, we have considered a version using the speaker embedder network to assign the target speaker and a version using oracle assignment of the target speaker. First, it can be observed that as expected the performance of the bidirectional one-step methods (SC-TSE) is larger than the performance of the unidirectional one-step methods. Second, both in unidirectional as well as bidirectional mode, the performance is significantly improved when using customized LSTM cells instead of standard LSTM cells. When customizing both forget and input gates, the mean SDR is improved by 0.99 dB (unidirectional) and 0.07 dB (bidirectional) compared to standard LSTM cell, while the mean SDR is only improved by 0.14 dB (unidirectional) and 0.01 dB (bidirectional) compared to only customizing the forget gate. The mean SDR is further improved when using the proposed customized auxiliary-gated LSTM cells, both in unidirectional mode (1.97 dB) and bidirectional mode (0.68 dB), compared to standard LSTM cells. It can also be observed that for each proposed variant of LSTM cells, the performance improvement when switching from unidirectional to bidirectional mode is not substantial, particularly for the customized auxiliary-gated LSTM. One possible reason is that the most crucial context information for the target speaker comes from the recent past, making future context less essential. Third, it can be observed that the performance of both two-step methods decreases when using the speaker embedder network to assign the target speaker compared to oracle assignment. The performance decreases by 0.5 dB for UPIT-LSTM and 0.6 dB for Conv-TasNet. Finally, it can be observed that both proposed one-step methods using variants of LSTM cells outperform the baseline two-step methods for target speaker extraction.

3.4.2 *Evaluation on 3-speaker mixtures*

In this Section, we investigate the effect of having more interfering speakers in the mixture during testing than during training. Since in this case two-step methods anyway do not perform well, we only consider one-step methods (SC-TSE). Table 3.3 shows the mean SDR of the same one-step methods as in Section 3.4.1 (trained on 2-speaker mixtures), but now evaluated on 3-speaker mixtures. As expected, it can be observed that the performance of all one-step methods systematically degrades compared to the results in Table 3.2. This can be explained by the fact that these systems are not aware of two interfering speakers at the time of training and hence fail to generalize for such conditions.

Systems	SDR (dB)	#Param
Mixture	0.14	-
One-step methods (unidirectional)		
Standard LSTM [15]	7.25	8.5 M
Customized LSTM (F)	8.10	7.8 M
Customized LSTM (F+I)	8.24	6.1 M
Customized auxiliary-gated LSTM	**9.22**	9.3 M
One-step methods (bidirectional)		
Standard BLSTM [15]	8.59	15.8 M
Customized BLSTM (F)	8.65	13.3 M
Customized BLSTM (F+I)	8.66	10.8 M
Customized auxiliary-gated BLSTM	**9.27**	16.8 M
Two-step methods (unidirectional)		
UPIT-LSTM [5]	7.74	8.6 M
Conv-TasNet [7]	7.82	5.0 M
Two-step methods - oracle assignment (unidirectional)		
UPIT-LSTM [5] - oracle	8.19	8.6 M
Conv-TasNet [7] - oracle	8.42	5.0 M

Table 3.2: Mean SDR and total number of parameters (#Param) of baseline and proposed one-step methods (SC-TSE) (unidirectional and bidirectional mode), baseline two-step methods (unidirectional mode), and baseline two-step methods with oracle assignment of target speaker (unidirectional mode). All systems are trained and evaluated on 2-speaker mixtures.

3.4.3 *Evaluation on multi-condition mixtures*

As shown in Table 3.3, the performance of SC-TSE systems degraded significantly due to the mismatch in training and testing conditions. In this section, we focus on improving the robustness of the considered SC-TSE systems against mismatch in training and testing conditions, in regards to the mixture signal.

In order to make the SC-TSE systems more robust against multiple interfering speakers and background noise, multi-condition mixture is utilized to train the systems. In this experiment, all considered systems are trained with 2-speaker mixtures, 3-speaker mixtures, 1-speaker mixtures with background noise (noisy 1-speaker mixtures) and 2-speaker mixtures with background noise (noisy 2-speaker mixtures).

Systems	SDR (dB)
Mixture	-3.05
One-step methods (unidirectional)	
Standard LSTM [15]	0.03
Customized LSTM (F)	0.12
Customized LSTM (F+I)	0.32
Customized auxiliary-gated LSTM	**0.66**
One-step methods (bidirectional)	
Standard BLSTM [15]	0.47
Customized BLSTM (F)	0.51
Customized BLSTM (F+I)	0.55
Customized auxiliary-gated BLSTM	**0.89**

Table 3.3: Mean SDR of baseline and proposed one-step methods (SC-TSE) (unidirectional and bidirectional mode). All systems are trained on only 2-speaker mixtures and evaluated on 3-speaker mixtures.

The performance is evaluated separately for 2-speaker mixtures, 3-speaker mixtures, noisy 1-speaker mixtures and noisy 2-speaker mixtures. To investigate the robustness of the systems, the performance is also evaluated for 3-speaker mixtures with background noise (noisy 3-speaker mixtures), which is not used during training.

Similarly as before, the speech samples for training are chosen from the 100 hours training set and official validation set of the Librispeech dataset [196], while the noise samples are chosen from the DNS challenge dataset [199]. To create noisy mixtures, the noise is added at an SNR which is randomly chosen between -5 and 10 dB. All together, we have generated 320 hours of training data and 80 hours of validation data, considering 2-speaker mixtures, 3-speaker mixtures, noisy 1-speaker mixtures and noisy 2-speaker mixtures in equal proportion. For testing, we have generated 35 hours of testing data considering 7 hours for each mixture type. The speech samples are chosen from the official test set of the Librispeech dataset, while the noise samples are chosen from the MUSAN dataset [200], which makes the test set completely disjoint from the training and validation sets.

Table 3.4 shows the mean SDR, SI-SDR, WB-PESQ and DNSMOS scores for 2-speaker mixtures and 3-speaker mixtures. When using multi-condition mixtures for training, the performance for 2-speaker mixtures is improved for the baseline as well as the proposed systems compared to training only on 2-speaker mixtures (see mean SDR results in Table 3.2). Although this is unexpected, similar results have been reported in [16]. When comparing the mean SDR results between Table 3.3 and Table 3.5 for 3-speaker mixtures, it can be observed that training with multi-condition

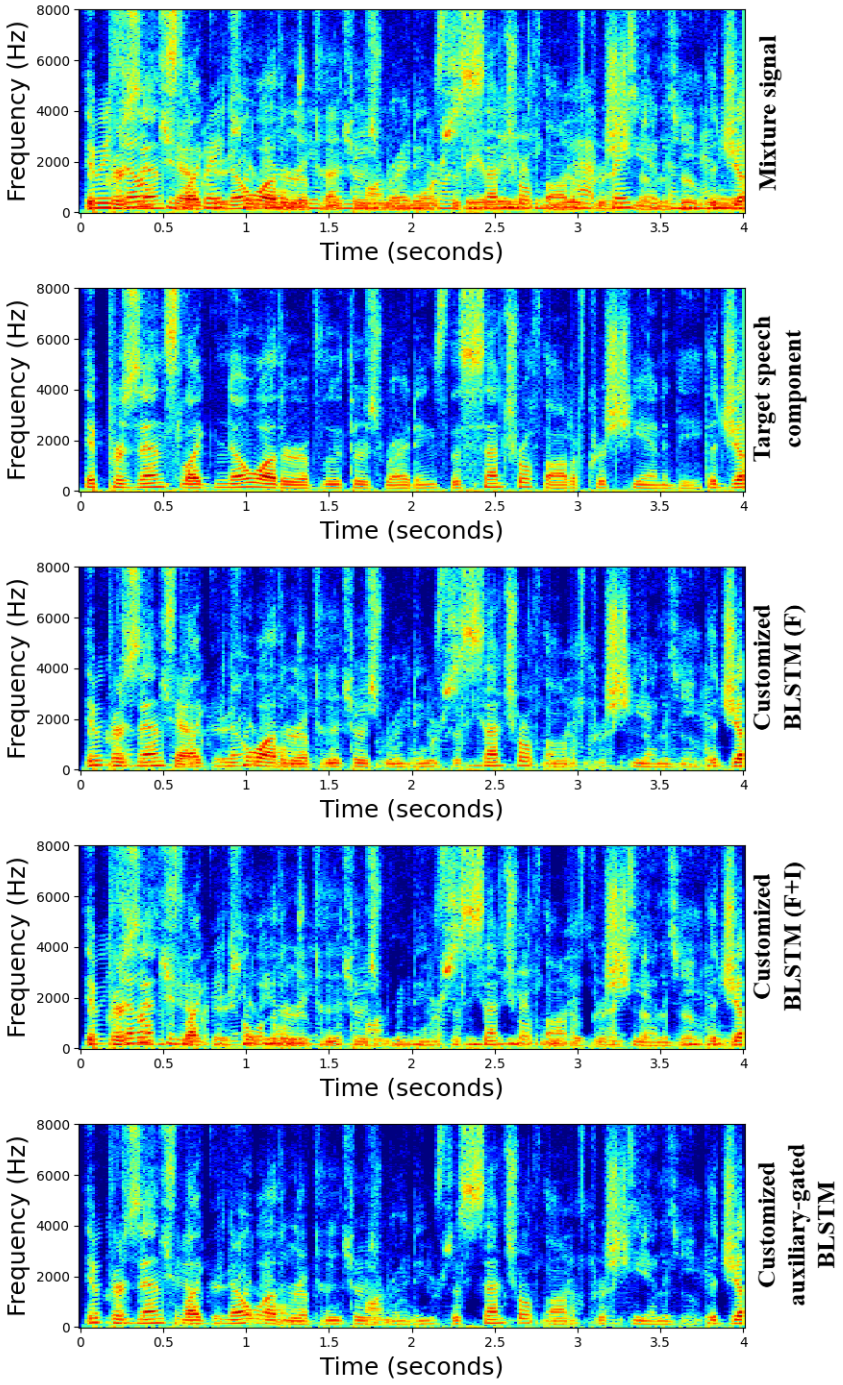

Fig. 3.10: Example spectrogram for the mixture signal, the target speech component, and the estimated target speech component using the customized BLSTM (F), the customized BLSTM (F+I), and the customized auxiliary-gated BLSTM.

Systems	2-speaker mixtures				3-speaker mixtures			
	SDR	SI-SDR	WB-PESQ	DNSMOS	SDR	SI-SDR	WB-PESQ	DNSMOS
Mixture	0.14	0.04	1.08	3.01	-3.05	-3.20	1.05	2.83
One-step methods (unidirectional)								
Standard LSTM [15]	8.01	7.47	1.43	3.15	2.81	2.19	1.14	2.89
Customized LSTM (F)	8.17	7.92	1.73	3.27	3.03	2.88	1.26	2.95
Customized LSTM (F+I)	8.22	7.99	1.74	3.27	3.11	2.93	1.26	2.95
Customized auxiliary-gated LSTM	**9.24**	**8.61**	**1.76**	**3.29**	**3.80**	**3.04**	**1.27**	**2.97**
One-step methods (bidirectional)								
Standard BLSTM [15]	8.68	8.50	1.70	3.26	3.09	2.87	1.24	2.97
Customized BLSTM (F)	8.72	8.65	1.74	3.27	3.17	3.00	1.26	2.99
Customized BLSTM (F+I)	8.72	8.66	1.74	3.27	3.17	3.01	1.26	2.99
Customized auxiliary-gated BLSTM	**9.41**	**8.85**	**1.77**	**3.30**	**3.96**	**3.21**	**1.28**	**3.01**

Table 3.4: Mean SDR, SI-SDR, WB-PESQ and DNSMOS of baseline and proposed one-step methods (SC-TSE) (unidirectional and bidirectional mode). All systems are trained on multi-condition mixtures and evaluated on 2-speaker mixtures and 3-speaker mixtures.

Systems	Noisy 2-speaker mixtures				Noisy 3-speaker mixtures			
	SDR	SI-SDR	WB-PESQ	DNSMOS	SDR	SI-SDR	WB-PESQ	DNSMOS
Mixture	-4.17	-4.35	1.04	2.55	-6.39	-6.66	1.04	2.50
One-step methods (unidirectional)								
Standard LSTM [15]	3.98	3.35	1.30	2.89	0.24	-0.49	1.09	2.75
Customized LSTM (F)	4.01	3.78	1.30	2.93	0.73	-0.14	1.13	2.80
Customized LSTM (F+I)	4.04	3.81	1.31	2.93	0.73	**-0.10**	1.13	2.81
Customized auxiliary-gated LSTM	**4.91**	**4.12**	**1.32**	**2.96**	**0.75**	-0.16	**1.14**	**2.83**
One-step methods (bidirectional)								
Standard BLSTM [15]	4.01	3.98	1.30	2.93	0.66	-0.14	1.11	2.80
Customized BLSTM (F)	4.19	4.09	1.31	2.93	0.74	-0.06	**1.13**	2.81
Customized BLSTM (F+I)	4.21	4.11	1.31	2.94	0.74	-0.07	**1.13**	2.81
Customized auxiliary-gated BLSTM	**5.23**	**4.48**	**1.33**	**2.98**	**0.80**	**-0.05**	**1.13**	**2.83**

Table 3.5: Mean SDR, SI-SDR, WB-PESQ and DNSMOS of baseline and proposed one-step methods (SC-TSE) (unidirectional and bidirectional mode). All systems are trained on multi-condition mixtures and evaluated on 2-speaker mixtures and 3-speaker mixtures with background noise.

Systems	Noisy 1-speaker mixtures			
	SDR	SI-SDR	WB-PESQ	DNSMOS
Mixture	2.26	2.23	1.26	2.74
One-step methods (unidirectional)				
Standard LSTM [15]	12.04	11.98	2.08	3.30
Customized LSTM (F)	12.54	12.03	2.22	3.31
Customized LSTM (F+I)	12.91	12.32	2.22	2.32
Customized auxiliary-gated LSTM	**13.66**	**13.06**	**2.25**	**3.34**
One-step methods (bidirectional)				
Standard BLSTM [15]	12.60	12.11	2.22	3.33
Customized BLSTM (F)	12.88	12.81	2.24	3.35
Customized BLSTM (F+I)	12.98	12.79	2.24	3.35
Customized auxiliary-gated BLSTM	**13.79**	**13.20**	**2.27**	**3.38**

Table 3.6: Mean SDR, SI-SDR, WB-PESQ and DNSMOS of baseline and proposed one-step methods (SC-TSE) (unidirectional and bidirectional mode). All systems are trained on multi-condition mixtures and evaluated on 1-speaker mixtures with background noise.

mixtures significantly improves the performance for all systems. For both 2-speaker as well as 3-speaker mixtures, all three proposed variants of LSTM cells show a consistent improvement in terms of all performance measures compared to standard LSTM cells. The best performance is obtained when using the customized auxiliary-gated LSTM cells, both in unidirectional as well as bidirectional mode. To offer a more intuitive comparison between the proposed variants, Fig. 3.10 shows example spectrograms of the mixture signal, the target speech component in the mixture, and the estimated target speech components using the customized BLSTM (F), the customized BLSTM (F+I) and the customized auxiliary-gated BLSTM. It can be observed that the spectrogram produced by the customized auxiliary-gated BLSTM more closely resembles the spectrogram of the target speech component, whereas the spectrograms produced by both customized BLSTM (F) and customized BLSTM (F+I) contain more residuals from the interfering speakers and have a similar structure.

Table 3.5 and Table 3.6 show the mean SDR, SI-SDR, WB-PESQ and DNSMOS scores for noisy 2-speaker mixtures, noisy 3-speaker mixtures, and noisy 1-speaker mixtures. Similar to Table 3.4, all proposed variants of LSTM cells improve the performance compared to standard LSTM cells. The performance improvement for noisy 1-speaker mixtures indicates that the proposed SC-TSE systems can also be utilized for personalized speech enhancement. Furthermore, the performance for noisy 3-speaker mixtures shows that training on multi-condition mixtures enables the systems to generalize to these mixtures, which were not included during training.

In conclusion, for all considered mixtures, the best performance is obtained when using proposed customized auxiliary-gated LSTM cells, both in unidirectional as well as bidirectional mode.

3.5 Summary

This chapter explored single-channel target speaker extraction in the time-frequency domain, focusing on optimizing the speaker embedder and speaker separator networks separately.

To enhance the speaker extraction performance of an standard LSTM-based baseline system, we proposed three novel variants of LSTM cells, specifically designed for the speaker separator network of SC-TSE system. The first variant focused on modifying only the forget gate, ensuring that it selectively retains relevant target speaker information while disregarding other sources present in the mixture. The second variant extended the first variant by also customizing the input gate, allowing better resetting of the cell state with target speaker information. The third variant introduced a novel auxiliary-modulation gate within the LSTM cell, which interacts with the forget and input gates to better capture long-term and short-term target speaker-specific features.

The proposed variants were evaluated on 2-speaker, 3-speaker, noisy 2-speakers, noisy 3-speakers, and noisy 1-speaker mixtures in both unidirectional and bidirectional modes, comparing their performance against a standard LSTM-based baseline SC-TSE system. Experimental results demonstrated that all proposed LSTM variants outperformed standard LSTM cells, with the auxiliary-gated LSTM achieving the best performance in both modes. Additionally, the proposed systems showed better generalization capabilities, even under challenging conditions such as noisy and unseen 3-speaker mixtures, when trained using multi-condition mixtures. Furthermore, all proposed variants of LSTM cells outperformed both considered two-step baseline methods based on BSS.

Future work could explore the potential benefit of applying customized LSTM variants with different modalities of auxiliary information for target speaker extraction as well as for other speech processing applications, such as own voice reconstruction in hearables exploiting an in-ear microphone.

4

CONFORMER-BASED ARCHITECTURES FOR SPEAKER-CONDITIONED TARGET SPEAKER EXTRACTION IN TIME DOMAIN

As an alternative to performing target speaker extraction in the time-frequency domain by estimating a real-valued mask, as in Chapter 3, this chapter focuses on extracting the target speaker in the time domain. In contrast to the separate training of the speaker embedder and separator networks in Chapter 3, both networks are jointly trained in an end-to-end learning process in this chapter (see Fig. 4.1).

Inspired by the success of the convolution-augmented transformer (conformer) [201], and its ability to capture both local and global context information, we propose two novel conformer-based architectures for the speaker separator network. The conformer is designed to model local-context dependencies through convolutional operations and global-context dependencies through MHSA. Both proposed architectures aim to perform target speaker extraction in the time domain. The first proposed architecture, Conformer-FFN, consists of stacked conformer blocks and external feed-forward blocks, aiming to capture both local and global context information efficiently while maintaining parameter efficiency. The second proposed architecture, TCN-Conformer, consists of stacked TCN blocks and conformer blocks, aiming to capture local-context information using TCN blocks first and then both local and global context information using conformer blocks. Traditional MHSA [202] is used in each conformer block for both proposed architectures.

This chapter evaluates the effectiveness of both proposed architectures against a TCN-based baseline system (SpEx+) [17]. While traditional MHSA [202] in each conformer block can efficiently model global-context dependencies, its high memory usage and computational complexity pose challenges for real-time applications. To

This chapter is partly based on the following publication:
[103] R. Sinha, M. Tammen, C. Rollwage, and S. Doclo, "Speaker-conditioning single-channel target speaker extraction using conformer-based architectures," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sept. 2022, pp. 1–5
[104] R. Sinha, C. Rollwage, and S. Doclo, "Real-time Single-channel Speaker-conditioned Target Speaker Extraction using TCN-Conformer with Efficient Self-attention Mechanisms," in *Proc. European Signal Processing Conference (EUSIPCO)*, Palermo, Italy, Sept. 2025, pp. 1–5

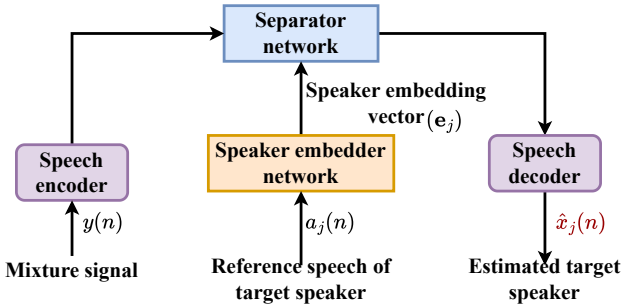

Fig. 4.1: Block diagram of SC-TSE system operating in the time domain.

address this, we further extend the TCN-Conformer architecture by replacing traditional MHSA with linear MHSA in each conformer block, making it more suitable for real-time target speaker extraction.

The remainder of this chapter is organized as follows: Section 4.1 provides a general overview of SC-TSE algorithms in the time domain (briefly revisiting Section 2.2.2.3). Section 4.2 discusses the speaker embedder network used in this chapter, followed by a detailed discussion of the proposed conformer-based speaker separator networks. Section 4.3 discusses the experimental setup, including the datasets, network architectures, the extended TCN-Conformer with linear MHSA for real-time target speaker extraction, and the training and evaluation hyperparameters. Section 4.4 presents the experimental results and performance analysis of the considered SC-TSE systems. Finally, Section 4.5 provides a summary of this chapter.

4.1 SC-TSE in time domain

In this section, we briefly formulate the target speaker extraction in the time domain, building on the definition provided in Section 2.2.2.3. Similar to [17], we consider a speaker separator network that aims at extracting the target speaker signal $x_j(n)$ by jointly optimizing the speaker embedder network and speaker separator network. The speaker embedder network not only estimates the target speaker embedding $\mathbf{e}_j$ from the reference speech $a_j(n)$ but simultaneously also contributes to the speaker extraction.

$$\mathbf{e}_j = \boldsymbol{\phi}^{emb}(a_j(n)), \tag{4.1}$$

As depicted in Fig. 4.1 (see also Section 2.2.2.3), the speech encoder transforms segments of the mixture signal into their latent representation (learned features), while the speech decoder reconstructs the target speaker signal from the masked encoded representation. The goal of the speaker separator network is to estimate the mask using the target speaker embedding $\mathbf{e}_j$ and the learned feature representation of the mixture signal.

$$\hat{x}_j(n) = \boldsymbol{\phi}^{sep}(y(n), \mathbf{e}_j), \tag{4.2}$$

where ϕ^{emb} and ϕ^{sep} denote the speaker embedder network and speaker separator network, respectively.

4.2 Overview of SC-TSE system architecture

In this section, we present an overview of the SC-TSE system architecture used in this chapter, followed by a detailed discussion of its individual components. Section 4.2.1 discusses the speaker embedder network, Section 4.2.2.1 discusses the multi-scale speech encoder and decoder, and Section 4.2.2.3 discusses the proposed speaker separator architectures, after briefly reviewing the baseline system in Section 4.2.2.2.

4.2.1 *Speaker embedder network*

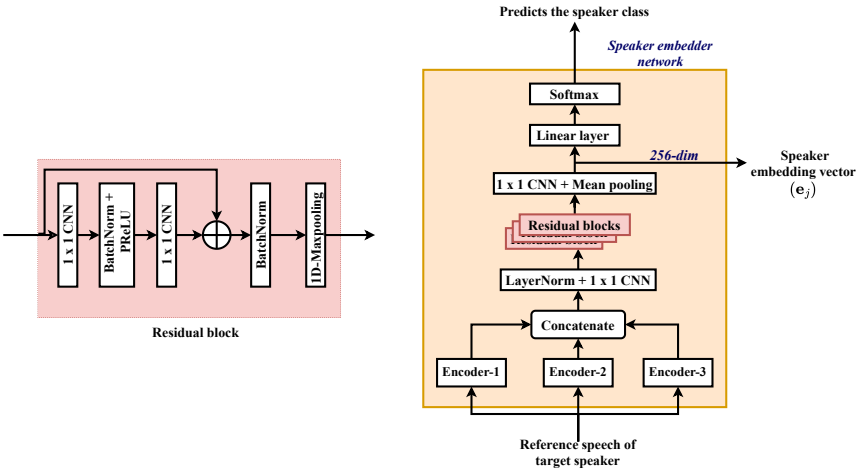

Fig. 4.2: Utilized ResNet-based speaker embedder network. The left side represents each residual block, while the right side represents the complete speaker identification system from which the target speaker embedding vector is obtained.

Similarly to [17], we have used the same ResNet-based speaker embedder network (see Fig. 4.2) to generate the target speaker embedding from the reference speech of the target speaker. The embedder network consists of a 1-D CNN layer that processes reference speech of the target speaker, followed by three residual blocks, which capture discriminative speaker characteristics. After the residual blocks, another 1-D CNN is utilized to project the resultant representations into a fixed-dimensional utterance-level speaker embedding vector. The final projection utilizes a mean pooling operation to aggregate the information across time. The rest of the layers after mean pooling contribute in classifying the speakers during the training. The speaker embedder network generates a fixed 256-dimensional embedding vector from the reference speech of the target speaker.

The input and output dimensions of the residual blocks are fixed to (256, 256), (256, 512), and (512, 512) respectively. Each residual block consists of two CNNs with a kernel size of 1 and a 1-D max-pooling layer. A batch normalization layer and parametric ReLU (PReLU) are used to normalize and apply non-linearity to the outputs obtained from each CNN layer. To improve gradient flow, a skip connection is utilized. This skip connection adds the input of the residual block to the output of the second batch-normalized layer. Additionally, a 1-D max-pooling layer with a kernel size of 1×3 is applied to discard silent regions and effectively reduce the feature dimensions by a factor of three, which enhances the quality of the speaker embedding by focusing on speech-relevant features.

4.2.2 *Speaker separator network*

The speaker separator network consists of three processing blocks: a speech encoder, a separator, and a speech decoder. In this section, we discuss each of these components in detail.

4.2.2.1 *Multi-scale speech encoder and decoder*

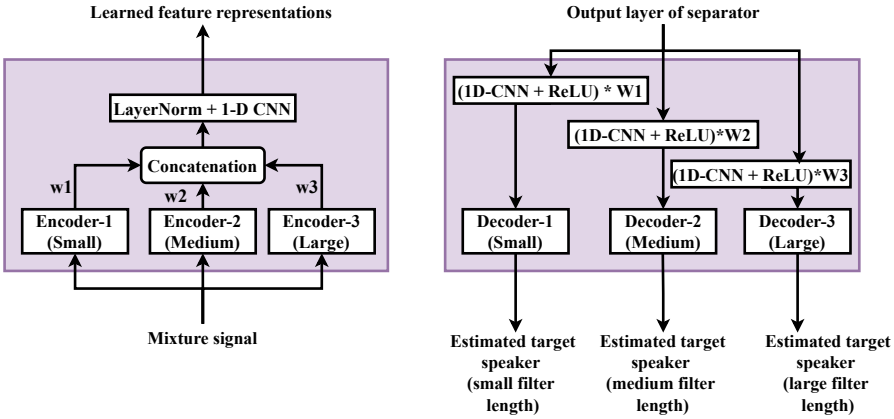

Fig. 4.3: Multi-scale speech encoder and decoder used for all considered SC-TSE systems.

We have used the same multi-scale speech encoder and decoder as the baseline system [17] (Fig. 4.3) to effectively capture features at different temporal resolutions. The encoder transforms the mixture signal into a unified latent space using multiple parallel 1-D CNN layers with varying filter lengths. While the number of temporal scales can be adjusted, we focus on three distinct levels: small, medium, and large. Small-scale filters (shorter filter lengths) capture fine-grained temporal details while maintaining computational efficiency, while medium and large-scale filters compen-

sate for the short-scale limitation by capturing broader contextual dependencies, ensuring a balance between local and global features. At first, the multi-scale feature representations of the mixture signal are obtained using the encoder, then the time-domain target speaker signals are reconstructed using the modulated responses at each scale, i.e., low, medium, and high using the speech decoder. The modulated responses are obtained by the element-wise multiplication of the mask estimated by the separator network and the representation of the mixture signal obtained using the speech encoder.

Throughout this chapter, we use the same multi-scale speech encoder and decoder with the same hyperparameters.

4.2.2.2 *Baseline separator network*

We consider a TCN-based baseline system [17], where both the speaker embedder and separator networks are jointly optimized to perform the speaker extraction in the time domain. The architecture is based on the stacking of TCN blocks consisting of two 1-D CNN layers, and two PReLU activation functions with global layer normalizations, along with a dilated depth-wise separable convolution (DDS-CNN) layer. The DDS-CNN layer incorporates an exponentially increasing dilation factor of 2^b, where b ranges from 0 to 7. Each stack consists of 8 TCN blocks, repeated for 4 times. At the start of each stack, the input to the very first TCN block is the concatenation of the target speaker embedding vector and the learned feature representations of the mixture signal obtained using the multi-scale speech encoder. To ensure the target speaker embedding is aligned with the features obtained from the encoder, the speaker embedding is repeated and concatenated along the feature dimension.

4.2.2.3 *Proposed separator networks*

In this section, we discuss the proposed conformer-based architectures (Conformer-FFN and TCN-Conformer) for the speaker separator network, after briefly reviewing the conformer block [201].

Conformer block

A conformer block (see Fig. 4.4(a)) consists of four sequentially stacked blocks: a feed-forward block (see Fig. 4.4(b)), a MHSA block (see Fig. 4.5(b)), a convolution block (see Fig. 4.5(a)), and again a second feed-forward block at the end. The MHSA block is enclosed between two feed-forward blocks along with a convolutional block. This enclosed structure is inspired by the system proposed in [203], which replaces the standard feed-forward block in the transformer with two half-step feed-forward layers, mainly one positioned before the MHSA block and the other after it. Each block of conformer is stacked in a structured way, enabling the conformer block to effectively model both local and global dependencies. Local dependencies captured through conformer block allow the SC-TSE system to capture fine-grained spectral and temporal patterns in the learned feature representation of the mixture signal obtained using the speech encoder that align with the target speaker characteris-

tics with the help of speaker embedding. In contrast, global dependencies allow the SC-TSE system to model long-range contextual information across the entire mixture signal, which is particularly important to distinguish the target speaker from interfering sources such as other speakers and background noise.

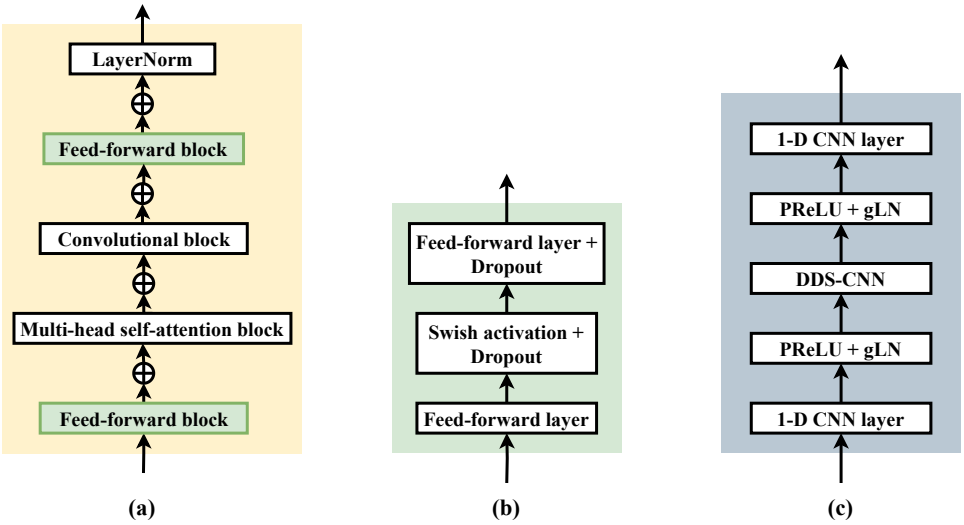

Fig. 4.4: (a) A conformer block, (b) a feed-forward block, and (c) a TCN block in the speaker separator network.

The output of the conformer block can be expressed as:

$$\tilde{\mathbf{y}}_{\text{se}} = \mathbf{y}_{\text{se}} + \frac{1}{2}FFN(\mathbf{y}_{\text{se}})$$

$$\mathbf{y}'_{\text{se}} = \tilde{\mathbf{y}}_{\text{se}} + MHSA(\tilde{\mathbf{y}}_{\text{se}})$$

$$\mathbf{y}''_{\text{se}} = \mathbf{y}'_{\text{se}} + Conv(\mathbf{y}'_{\text{se}})$$

$$\mathbf{y}_{\text{conf}} = \text{LayerNorm}(\mathbf{y}''_{\text{se}} + \frac{1}{2}FFN(\mathbf{y}''_{\text{se}})), \tag{4.3}$$

where $\mathbf{y}_{\text{se}}$ and $\mathbf{y}_{\text{conf}}$ denote the input to the first feed-forward block of the conformer block and output of the conformer block, respectively. $\tilde{\mathbf{y}}_{\text{se}}$ denotes the input to the MHSA block, and $\mathbf{y}'_{\text{se}}$ denotes the input to the convolutional block.

Proposed Conformer-FFN architecture

Fig. 4.6 depicts the proposed Conformer-FFN architecture which is based on stacking of conformer blocks and external feed-forward blocks. The motivation behind

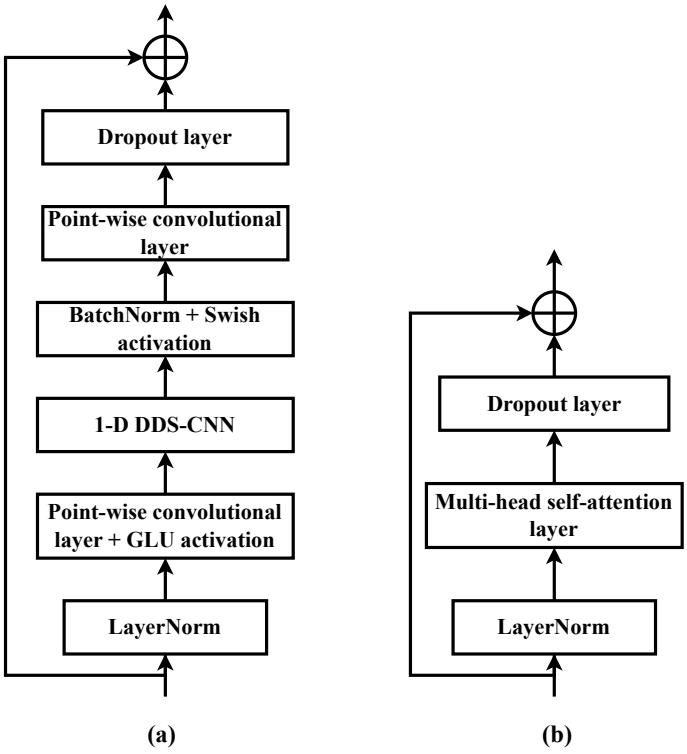

Fig. 4.5: (a) The convolutional block and (b) the MHSA block used in the conformer block.

this architecture is to utilize both local and global context features simultaneously using conformer blocks. The arrangement of components within each conformer block follows the structure depicted in Fig. 4.4(a). It should be noted that each conformer block consists of traditional MHSA. In our proposed architecture, each conformer block is followed by an external feed-forward block. The external feed-forward block consists of two feed-forward layers, Swish activation [204], and dropout layers (see Fig. 4.4(b)), the same as the feed-forward block used inside the conformer block. The output dimension of the external feed-forward block is half of the input dimension, leading to two advantages. First, we obtain a fixed dimensional input for all conformer blocks. Second, the overall number of parameters for the speaker separator network is reduced. The input to the first conformer block is the concatenation of the learned features obtained from the mixture signal using the multi-scale speech encoder and the target speaker embedding, while the inputs to the rest of the conformer blocks are the concatenation of the external feed-forward block output and the target speaker embedding. Similarly to the considered baseline system [17] the target speaker embedding is repeated and concatenated along the feature dimension.

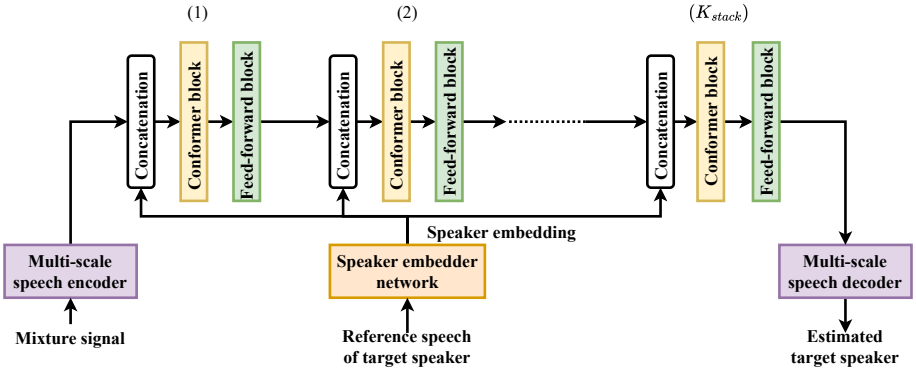

Fig. 4.6: Proposed Conformer-FFN architecture, where K_{stack} denotes the number of stacks of conformer and feed-forward blocks.

Proposed TCN-Conformer architecture

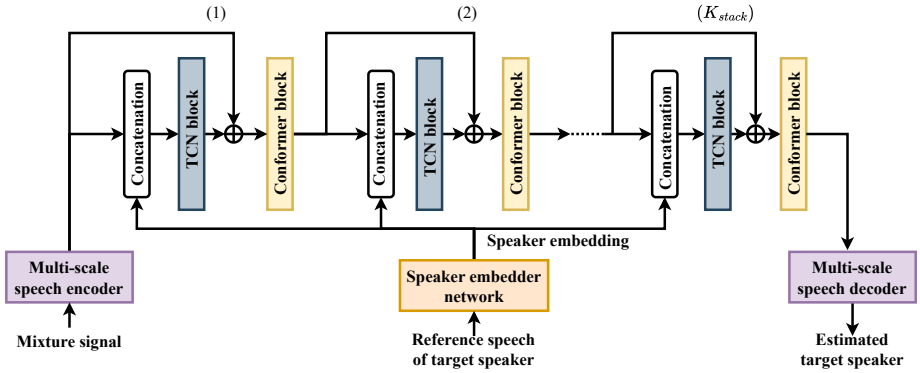

Fig. 4.7: Proposed TCN-Conformer architecture, where K_{stack} denotes the number of stacks of TCN and conformer blocks.

Fig. 4.7 depicts the proposed TCN-Conformer architecture, which is based on the stacking of TCN blocks and conformer blocks, i.e., each TCN block is followed by a conformer block. The motivation behind this architecture is to first utilize local context features using TCN blocks and then exploit both local and global context features using conformer blocks. The arrangement of components within each conformer block follows the structure depicted in Fig. 4.4(a). It should be noted that each conformer block consists of traditional MHSA. Each TCN block [7] (see Fig. 4.4(c)) consists of two 1-dimensional CNN (1D-CNN) layers, two PReLU activations with global layer normalization (gln), and one dilated depth-wise separable CNN layer (DDS-CNN). The conformer block consists of the same architecture as discussed in Section 4.2.2.3. The input to the first TCN block is the concatenation

of the learned features obtained from the mixture signal using the multi-scale speech encoder and the target speaker embedding estimated using the speaker embedder network, while the inputs to the rest of TCN blocks are the concatenation of the conformer block output and the target speaker embedding. Similarly to the TCN-based baseline system [17] and the proposed Conformer-FFN system, the target speaker embedding is repeated and concatenated along the feature dimension.

4.3 Experimental setup

In this section, we discuss the experimental setup used for all considered SC-TSE systems. In Section 4.3.1, we discuss the datasets used for training and testing. In Section 4.3.2, we discuss the training setup and hyperparameters for both the speaker embedder and separator networks. In Section 4.3.3, we discuss the extended TCN-Conformer with linear MHSA for real-time target speaker extraction.

4.3.1 *Dataset*

To generate the training, validation, and test data, we have used two different datasets at a 16 kHz sampling rate: the Wall Street Journal (WSJ) dataset [137] for speech and the WHAM dataset [205] for background noise. The WSJ0 dataset consists of three subsets, namely: si_tr_s, si_dt_05, and si_et_05. The training and validation sets are created from the si_tr_s subset, while the test set is created using the si_dt_05 and si_et_05 subsets, ensuring that the test set contains entirely different speakers than those in training and validation sets. In total, the training set includes 101 speakers, the validation set 18 speakers, and the test set 16 speakers, maintaining a balanced distribution across genders and speaker identities. The WHAM dataset consists of the real-world background noise recorded in various indoor and outdoor environments. This dataset includes its own training, validation, and test splits.

Using these datasets, we generate three types of speech mixtures: 2-speaker mixtures, 3-speaker mixtures, and 2-speaker mixtures with background noise (noisy 2-speaker mixtures). The 2-speaker mixtures are generated by randomly selecting two speakers and mixing them at an SNR between 0 and 5 dB, where the first speaker is considered the target speaker, and the second as the interfering speaker. To obtain the target speaker embedding, a different utterance from the target speaker is used as the reference speech. Similarly, for the 3-speaker mixtures, two interfering speakers are randomly chosen. After mixing both interfering speakers with the same power, the resulting mixture of interfering speakers is mixed with the target speaker at an SNR between 0 dB and 5 dB. For the noisy 2-speaker mixtures, the target and interfering speakers are selected from the WSJ0 dataset, and the background noise is selected from the training, validation, and test sets of the WHAM dataset. These noisy mixtures are generated following the official WHAM dataset simulation scripts for 2-speaker mixtures. In total, we generate 47926 utterances for training, 12792 for validation, and 7478 for testing, considering all three types of mixtures.

4.3.2 *Training setup*

We have used the same hyperparameters for the speaker embedder network in all considered SC-TSE systems. Each TCN block used in the proposed TCN-Conformer system uses the hyperparameter settings as the first TCN block utilized in the baseline system [17]. The input and convolutional size of the TCN block are fixed to 512, and the kernel size is fixed to 3. Each conformer block utilizes 8-head attention, while the convolutional kernel size is fixed to 31. The output of the first CNN layer is expanded with a factor of 3 in each block, while the output of the feed-forward layer is set to be 4 times the input size. The external feed-forward blocks in the proposed Conformer-FFN system (see Fig. 4.6) utilize an input dimension of 512 and an output dimension of 256, as well as a Swish activation function (same as the feed-forward block utilized in the conformer block).

To train all considered SC-TSE systems, we have used the weighted combination of multi-scale SI-SDR loss function for the speaker separator network and the CE loss function for the speaker embedder network, same as in [17]. The overall loss function is defined as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{MS-SI-SDR}}\big(x_j, \hat{x}_j\big) + \gamma\, \mathcal{L}_{\text{CE}}\big(\mathbf{C}, \hat{\mathbf{p}}\big), \tag{4.4}$$

where $\mathcal{L}_{\text{MS-SI-SDR}}$ is defined as:

$$\begin{aligned}\mathcal{L}_{\text{MS-SI-SDR}}\big(x_j, \hat{x}_j\big) = \big[&(1-\beta_1-\beta_2)\, \mathcal{L}_{\text{SI-SDR}}\big(x_j, \hat{x}_{js}\big) \\ &+ \beta_1\, \mathcal{L}_{\text{SI-SDR}}\big(x_j, \hat{x}_{jm}\big) \\ &+ \beta_2\, \mathcal{L}_{\text{SI-SDR}}\big(x_j, \hat{x}_{jl}\big)\big],\end{aligned} \tag{4.5}$$

where $\hat{x}_{js}$, $\hat{x}_{jm}$, and $\hat{x}_{jl}$ represent the estimated target speaker signals using the small, medium, and large filter, respectively. $\beta_1, \beta_2 \in [0, 1]$ and $\beta_1 + \beta_2 \leq 1$, and hence these weighting hyperparameters sum to one. $\mathcal{L}_{\text{SI-SDR}}$ represents the SI-SDR loss function as defined in (2.22). The CE loss measures the difference between the one-hot ground-truth vector $\mathbf{C}$ representing the true class labels for the speaker and the predicted speaker posterior vector $\hat{\mathbf{p}}$ using the speaker embedder network. γ represents the scaling factor. The CE loss is defined as:

$$\mathcal{L}_{\text{CE}}\big(\mathbf{C}, \hat{\mathbf{p}}\big) = -\sum_{i=1}^{I_c} C_i \log \hat{p}_i, \tag{4.6}$$

where $\hat{\mathbf{p}} = \text{softmax}(\mathbf{W}_{\text{emb}}\mathbf{e}_j)$, $\mathbf{W}_{\text{emb}}$ represents the weight matrix that maps the embedding vector to class scores, and I_c represents the number of speaker classes used for the speaker embedder network. A visual representation of estimated target speaker embeddings from the joint learning of the speaker embedder and separator network is shown in Fig. 4.8. To plot these embeddings, 256-dimensional embedding

vectors are estimated from the last layer of the embedder network before predicting the class probabilities and reducing them to two dimensions using t-SNE.

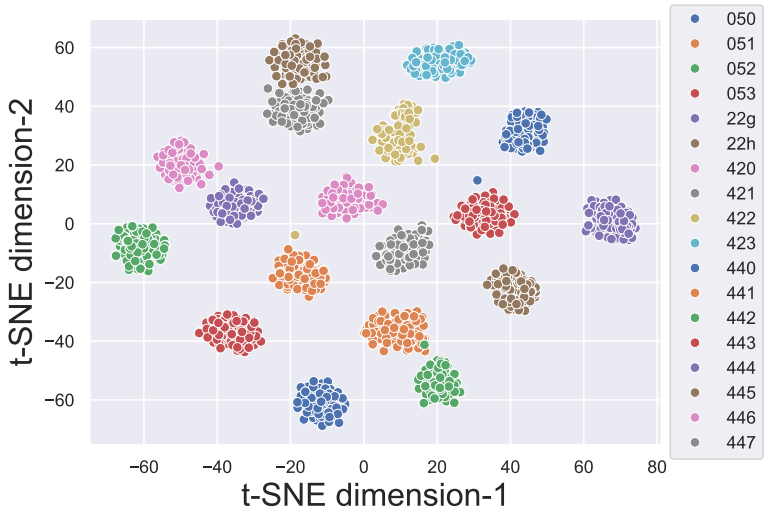

Fig. 4.8: Target speaker embeddings estimated on the reference speech of the target speaker from the test set utilizing the speaker embedder network of the jointly optimized SC-TSE system. To represent the speaker embeddings, t-SNE is used to reduce them in two dimensions. Each point on the plot represents one utterance of the target speaker, where colors represent the corresponding speakers.

All systems in this chapter are trained using the ADAM optimizer [198] with a learning rate of 0.001, employing a weighted combination of this multi-scale SI-SDR loss and the cross-entropy loss [17] across all mixture types. The weighting hyperparameters, β_1 and β_2 are set as 0.1, while the scaling factor for both loss functions, γ is set to 10 for each system. Each considered system is trained for 4-s of segments of audio signals for 150 epochs with an early stopping criterion of 6 epochs. The three filter lengths for the multi-scale encoder and decoder are set as 2.5 ms, 10 ms, and 20 ms for all considered systems, while the number of stacking for the proposed Conformer-FFN and TCN-Conformer is chosen as $K_{\text{stack}} \in \{1, 3, 4\}$. In the following section, we discuss the extension of the proposed TCN-Conformer system with $K_{\text{stack}} = 4$ for real-time target speaker extraction

4.3.3 *TCN-Conformer system for real-time target speaker extraction*

The traditional self-attention (SA) mechanism [202] as utilized in both proposed architectures has received significant attention for target speaker extraction [156], [206] due to its ability to capture complex dependencies, support parallel processing, and its flexibility in handling diverse input features. Despite its impressive performance, traditional SA often comes with high memory and computational costs [202]. The memory and computational costs of traditional SA scale quadratically

with the length of the mixture signal, making it resource-intensive and challenging for real-time applications. While earlier in Section 4.2.2.3, we explored the benefits of utilizing conformers with FFN and TCN for target speaker extraction, where each conformer block utilized the traditional MHSA. In this section, we focus on reducing both the memory and computational costs of our best performing proposed TCN-Conformer system (with $K_{\text{stack}} = 4$) to make it more suitable for real-time applications (see Fig. 4.9). We introduce two key modifications. First, we replace the traditional MHSA in each conformer block with linear MHSA [207], which scales linearly with input length. Second, we explore several system variants by progressively reducing the total number of parameters by a factor of 2 (Medium), 4 (Small), and 8 (XSmall). To evaluate the effectiveness of the proposed TCN-Conformer with linear MHSA, we compare it against traditional MHSA and a memory-efficient MHSA [208].

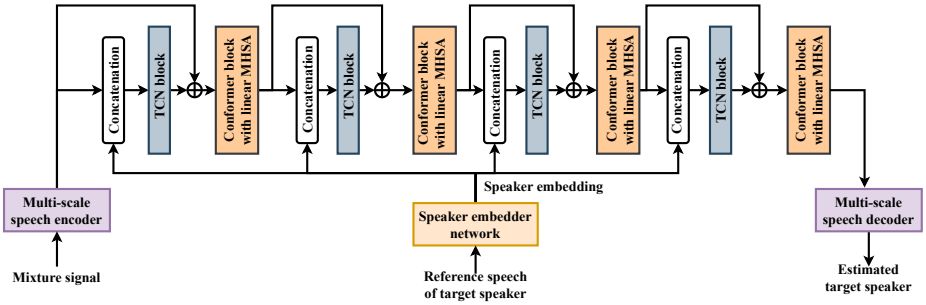

Fig. 4.9: Proposed TCN-Conformer system with linear MHSA in the conformer block.

In this section, we first review the traditional MHSA used in the TCN-Conformer system along with memory-efficient MHSA [208] (see 4.3.3.1), and then discuss the linear MHSA (see Section 4.3.3.2). We also discuss several system variants and their hyperparameters obtained through progressive reduction of the total number of parameters (see Section 4.3.3.3).

4.3.3.1 *TCN-Conformer system with traditional MHSA*

Each conformer block of the TCN-Conformer system (see Fig. 4.7) utilizes a traditional MHSA. For the simplicity of notation indexing, we represent input $\tilde{\mathbf{y}}_{\text{se}}$ to the MHSA block from (4.3) as $\tilde{\mathbf{y}}$ in this section. The traditional SA mechanism transforms each input feature vector $\tilde{\mathbf{y}}_{\text{i}} \in \mathbb{R}^d$ with dimension d into three vectors: query ($\mathbf{q}_{\text{i}} \in \mathbb{R}^{d_k}$), key ($\mathbf{k}_{\text{i}} \in \mathbb{R}^{d_k}$), and value ($\mathbf{v}_{\text{i}} \in \mathbb{R}^{d_v}$). The query and key have the same feature dimension d_k, while d_v denotes the feature dimension of the value. The similarity between the i-th query and the j-th key is computed as softmax $\left(\mathbf{q}_{\text{i}}^T \mathbf{k}_j\right)$. SA focuses on finding similarities between all pairs of positions, where for n positions, the queries, keys and values are represented as matrices $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$, $\mathbf{K} \in \mathbb{R}^{n \times d_k}$, and $\mathbf{V} \in \mathbb{R}^{n \times d_v}$, respectively. The output of a traditional SA is given as:

$$Att(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\mathbf{Q}\mathbf{K}^T\right)\mathbf{V}, \tag{4.7}$$

where softmax denotes the softmax operation. Since traditional SA issues a separate query for each position, it exhibits overall memory and computational costs of $O(n^2)$. MHSA consists of the concatenation of several parallel SA layers.

A memory-efficient MHSA mechanism has been proposed in [208], which closely preserves the same mathematical equivalence of traditional MHSA while optimizing memory usage. Unlike traditional MHSA, which stores the full attention matrix in (4.7), memory-efficient MHSA dynamically recomputes the attention matrix during backpropagation instead of retaining it in memory, which significantly reduces memory consumption while maintaining the accuracy and functionality of traditional MHSA. We use the memory-efficient MHSA in each conformer block of the TCN-Conformer system and compare the performance against traditional MHSA and proposed TCN-Conformer system with linear MHSA.

4.3.3.2 *TCN-Conformer system with linear MHSA*

The proposed TCN-Conformer system with linear MHSA (linear TCN-Conformer) uses the same block diagram as the TCN-Conformer system, where the traditional MHSA in each conformer block is replaced with linear MHSA (see Fig. 4.9) [207] to reduce both memory and computational costs. Similar to traditional SA, linear SA also transforms input features into queries, keys, and values through linear transformations. However, instead of treating the keys as n feature vectors in $\mathbb{R}^{d_k}$, linear SA interprets them as d_k feature maps [207]. Each feature map acts as a weight across all positions and aggregates the corresponding values through a weighted sum. The output of this linear SA layer is given as:

$$Att_{\text{lin}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}_q\left(\mathbf{Q}\right)\left(\text{softmax}_k(\mathbf{K})^T\mathbf{V}\right), \tag{4.8}$$

where softmax_q and softmax_k denote row-wise and column-wise softmax operations, respectively. Although these softmax operations on the query and key matrices ($\mathbf{Q}$, $\mathbf{K}$) are not the same as performing a single softmax on $\mathbf{Q}\mathbf{K}^T$ as in (4.7), they may closely approximate the overall effect. The property of $\text{softmax}(\mathbf{Q}\mathbf{K}^T)$ is that each row sums to 1, representing a normalized attention distribution over all positions. The matrix $\text{softmax}_q(\mathbf{Q})\text{softmax}_k(\mathbf{K})^T$ retains this property. Therefore, the linear SA mechanism in (4.8) offers a close approximation to traditional SA in (4.7) while significantly reducing memory and computational requirements, with both scaling linearly with input length.

4.3.3.3 *Proposed system variants*

First, we train the TCN-Conformer system with traditional and memory-efficient MHSA, and the proposed linear TCN-Conformer system in causal mode, ensuring that all systems have a similar number of parameters as the TCN-Conformer system with traditional MHSA (non-causal mode). For all systems, we consider 4 different

variants (see Table 4.1). A large variant with about 12.8 M parameters, a medium variant (factor 2 reduction), a small variant (factor 4 reduction), and an extra small variant (factor 8 reduction). To achieve the different variants, we varied the number of filters in the multi-scale speech encoder, the dimension of the MHSA in the conformer block, and the number of filters in the DDS-CNN layers, while keeping all other (hyper)parameters the same. For each variant, 4 stacks of TCN and conformer blocks are used in the separator network.

Variants	**Filters (speech encoder)**	**Attention-dimension**	**Filters (DDS-CNN)**
Large	256	256	512
Medium	1024	64	920
Small	512	64	512
XSmall	256	64	256

Table 4.1: (Hyper)parameter settings for the different variants of the TCN-Conformer system with traditional and memory-efficient MHSA, and proposed linear TCN-Conformer system.

4.4 Results and discussion

The experimental results are presented in two sections:

1. Comparison with baseline: The first Section 4.4.1 compares the performance of the proposed Conformer-FFN and TCN-Conformer systems against the TCN-based baseline system [17].
2. Performance analysis of TCN-Conformer system with linear MHSA: The second Section 4.4.2 compares the TCN-Conformer system with linear MHSA against TCN-Conformer systems with traditional and memory-efficient MHSA. This comparison highlights differences in both target speaker extraction performance and real-time processing capabilities.

4.4.1 *Comparison with baseline*

In this section, we compare the performance of each proposed Conformer-FFN system and TCN-Conformer system with the considered TCN-based baseline system [17]. We have utilized the SI-SDR as discussed in Section 2.4.1 as a performance measure.

In Section 4.4.1.1, both the baseline and the proposed systems are trained exclusively on 2-speaker mixtures. For the Conformer-FFN and TCN-Conformer-based systems, the number of stacking is set to $K_{\text{stack}} = 3$. Each system is then evaluated separately on three test sets: 2-speaker mixtures (2-mix), 3-speaker mixtures (3-mix), and noisy 2-speaker mixtures (noisy-mix). In Section 4.4.1.2, the baseline and the proposed systems are trained on multi-condition mixtures, which include 2-mix, 3-mix, and noisy-mix mixtures combined. Under this setting, each proposed system is trained with three different stacking configurations: $K_{\text{stack}} = 1$, $K_{\text{stack}} = 3$, and $K_{\text{stack}} = 4$.

Similar to the previous section, each system is evaluated separately on the test sets of 2-mix, 3-mix, and noisy-mix.

4.4.1.1 *Evaluation on 2-speaker mixtures*

Table 4.2 shows the mean SI-SDR scores for all considered SC-TSE systems trained exclusively on the 2-mix data, evaluating their effectiveness on 2-mix, 3-mix, and noisy-mix test sets. It can be observed that all considered SC-TSE systems significantly improve speaker extraction performance compared to the input mixtures, confirming their effectiveness in extracting the target speaker from overlapping input mixture. Among all considered systems, the proposed TCN-Conformer system achieves the best overall performance, outperforming both the baseline system and the proposed Conformer-FFN system across all mixture types. Specifically, the proposed TCN-Conformer system outperforms the baseline by 0.70 dB on 2-mix, 0.38 dB on 3-mix, and 2.06 dB on noisy-mix. In contrast, the proposed Conformer-FFN system underperforms compared to the baseline system for all mixture types. While its performance is slightly lower than the baseline system for 2-mix and 3-mix, it shows a significant performance drop on noisy-mix, showing a high sensitivity to the background noise. Additionally, it can be observed that all considered systems trained only with the 2-mix data achieve high performance for the 2-mix test set, but perform poorly when either an additional interfering speaker (3-mix) or additional background noise (noisy-mix) is added in the input mixture.

Systems	$K_{\textbf{stack}}$	**2-mix**	**3-mix**	**Noisy-mix**
Input mixture	-	2.51	-1.27	-3.21
Baseline system [17]	-	16.15	4.18	-2.30
Conformer-FFN	3	15.60	4.08	-3.64
TCN-Conformer	3	**16.85**	**4.56**	**-0.24**

Table 4.2: SI-SDR (dB) for the input mixture, baseline system, and proposed Conformer-FFN and TCN-Conformer systems trained only with the 2-speaker mixtures (2-mix).

4.4.1.2 *Evaluation on multi-condition mixtures*

Table 4.3 shows the mean SI-SDR (dB) scores for all considered SC-TSE systems trained on multi-condition mixtures, evaluating their effectiveness on 2-mix, 3-mix, and noisy-mix test sets. As observed in Section 4.4.1.1, all systems significantly outperform the input mixture, showing their ability to extract the target speaker from a mixture of overlapping speech. Additionally, both proposed Conformer-FFN and TCN-Conformer systems show performance improvements as the number of stacking between conformer and feed-forward blocks for the Conformer-FFN, and between TCN and conformer blocks for the TCN-Conformer increases. For the proposed Conformer-FFN system, increasing K_{stack} from 1 to 4 leads to a steady

improvement across all mixture types. However, despite this consistent gain, the best-performing Conformer-FFN system ($K_{\text{stack}} = 4$) still underperforms the baseline system, suggesting that deeper stacking alone is not sufficient. In contrast, the proposed TCN-Conformer system consistently outperforms Conformer-FFN across all stacking, indicating its architectural advantage. The performance difference between the proposed systems becomes more significant at $K_{\text{stack}} = 3$, where TCN-Conformer outperforms the Conformer-FFN by 2.44 dB (2-mix), 2.12 dB (3-mix), and 1.79 dB (noisy-mix). At $K_{\text{stack}} = 4$, TCN-Conformer achieves the best performance, outperforming each proposed Conformer-FFN system at every K_{stack}.

Systems	K_{stack}	**2-mix**	**3-mix**	**Noisy-mix**
Input mixture	-	2.51	-1.27	-3.21
Baseline system [17]	-	14.87	8.43	7.92
Conformer-FFN	1	11.99	6.34	6.30
Conformer-FFN	3	13.03	7.09	7.08
Conformer-FFN	4	14.07	7.67	7.56
TCN-Conformer	1	12.34	7.12	6.85
TCN-Conformer	3	**15.47**	**9.21**	**8.87**
TCN-Conformer	4	**17.51**	**10.70**	**9.32**

Table 4.3: SI-SDR (dB) for the input mixture, baseline system and proposed Conformer-FFN and TCN-Conformer systems trained on multi-condition mixtures (2-mix, 3-mix, and noisy-mix dataset).

While comparing the proposed TCN-Conformer ($K_{\text{stack}} = 4$) to the baseline, it improves the speaker extraction performance by 2.64 dB (2-mix), 2.27 dB (3-mix), and 1.40 dB (noisy-mix), showing its effectiveness in speaker extraction across all mixture types. Furthermore, a comparative analysis of Table 4.2 and Table 4.3 shows the impact of multi-condition training, which contributes to a more robust system. All systems achieve their best performance on 2-mix test set, but performance degrades when an additional interfering speaker or background noise is added. 3-mix test set condition presents a greater challenge due to increased speech overlap, leading to reduced SI-SDR scores across all systems. Noisy-mix condition further reduces the performance, particularly for weaker systems. Nevertheless, TCN-Conformer ($K_{\text{stack}} = 4$) remains the most robust, outperforming the baseline by 1.40 dB and the proposed Conformer-FFN ($K_{\text{stack}} = 4$) by 1.76 dB for noisy-mix test set.

Our best-performing TCN-Conformer-based system (with $K_{\text{stack}} = 4$) is further extended and evaluated for its real-time processing capabilities in the next section 4.4.2.

Variants	Systems	MHSA	Mode	2-mix	3-mix	noisy-mix	RTF	MACs	#Param
-	Input mixture	-	-	2.5	-1.3	-3.2	-	-	-
Large	TCN-Conformer	Traditional	Non-causal	17.5	10.7	9.3	-	17.92 G	12.8 M
	TCN-Conformer	Traditional	Causal	12.6	7.1	6.0	2.31	16.72 G	12.6 M
	TCN-Conformer	Memory-efficient	Causal	11.8	6.6	5.8	2.00	16.66 G	12.3 M
	Linear TCN-Conformer	Linear	Causal	**12.9**	**7.3**	**6.7**	**1.60**	14.05 G	12.3 M
Medium	TCN-Conformer	Traditional	Non-causal	15.8	9.5	8.7	-	15.24 G	6.4 M
	TCN-Conformer	Traditional	Causal	11.2	6.4	**6.1**	1.83	13.09 G	6.4 M
	TCN-Conformer	Memory-efficient	Causal	11.0	6.1	5.7	1.62	13.00 G	6.3 M
	Linear TCN-Conformer	Linear	Causal	**11.7**	**6.6**	5.9	**0.23**	9.27 G	6.3 M
Small	TCN-Conformer	Traditional	Non-causal	14.6	8.6	8.1	-	11.74 G	3.1 M
	TCN-Conformer	Traditional	Causal	10.9	6.3	5.9	1.41	9.27 G	3.1 M
	TCN-Conformer	Memory-efficient	Causal	10.7	6.0	5.7	1.07	9.19 G	3.0 M
	Linear TCN-Conformer	Linear	Causal	**11.4**	**6.4**	**6.1**	**0.12**	5.03 G	3.0 M
XSmall	TCN-Conformer	Traditional	Non-causal	14.0	8.0	7.9	-	8.92 G	1.7 M
	TCN-Conformer	Traditional	Causal	10.6	6.1	5.2	0.90	7.71 G	1.7 M
	TCN-Conformer	Memory-efficient	Causal	10.2	6.0	5.2	0.72	7.66 G	1.6 M
	Linear TCN-Conformer	Linear	Causal	**11.3**	**6.3**	**5.9**	**0.08**	3.31 G	1.6 M

Table 4.4: Mean SI-SDR (dB), real-time factor (RTF), total number of MACs per second, and number of parameters for different variants of TCN-Conformer systems with traditional and memory-efficient MHSA, and proposed linear TCN-Conformer systems.

4.4.2 *Performance analysis of TCN-Conformer system with linear MHSA*

Similarly to Section 4.4.1.2, the performance of all considered systems is evaluated separately on the 2-mix, 3-mix, and noisy-mix test sets. Besides using the SI-SDR as performance measures for target speaker extraction, we have also considered the computational and memory costs measured by the multiply-accumulate operations per second (MACS), the real-time factor (RTF), and the total number of parameters (#Param). MACs and #Param have been computed using the `Torchinfo` library of PyTorch. The RTF has been measured on an Intel Core i7-10850H CPU (2.7 GHz) as the time required to process an audio signal divided by its duration, where we have conducted 100 passes with 4-s segments of audio signals.

Table 4.4 shows the mean SI-SDR, computational cost, and memory usage for the TCN-Conformer system with traditional MHSA (evaluated in both non-causal and causal modes), memory-efficient MHSA (causal mode), and linear MHSA (causal mode). The results show that all considered systems significantly improve SI-SDR compared to the input mixtures. As expected, the TCN-Conformer system with traditional MHSA in causal mode shows a significant performance reduction compared to its non-causal mode. The TCN-Conformer system with memory-efficient MHSA shows slight reduction in RTF compared to the traditional MHSA, but no improvement in the speaker extraction performance. Notably, the proposed TCN-Conformer with linear MHSA outperforms TCN-Conformer system with both traditional and memory-efficient MHSA in terms of SI-SDR with a significant reduction in RTF across all variants and mixture types, except for the Medium variant in noisy-mix. Furthermore, while the number of MAC operations per second remains similar between the traditional and memory-efficient MHSA, the TCN-Conformer with linear MHSA achieves a significant reduction in number of MACs. Also, TCN-Conformer system with linear MHSA improves speaker extraction performance compared to both the traditional and memory-efficient MHSA. One possible reason is that linear MHSA generates smoother, more globally coherent temporal attention weights that suppress artifacts more effectively than traditional or memory-efficient MHSA. In terms of real-time processing capability, except for the XSmall TCN-Conformer systems, none of the other systems utilizing traditional or memory-efficient MHSA are real-time capable ($RTF > 1$). In contrast, the proposed TCN-Conformer system with linear MHSA is suitable for real-time processing across all variants except the Large variant (12.3 M parameters). For instance, compared to the XSmall variant of TCN-Conformer system with traditional MHSA, the TCN-Conformer with linear MHSA (1.6 M parameters, XSmall variant) improves SI-SDR by 0.7 dB for 2-mix, 0.2 dB for 3-mix, and 0.7 dB for noisy-mix, while achieving approximately 91% reduction in RTF. Furthermore, it can be observed that RTF reduction becomes more pronounced in smaller variants.

4.5 Summary

This chapter explored single-channel target speaker extraction in the time domain, focusing on optimizing the speaker embedder and speaker separator networks jointly.

First, we proposed two novel conformer-based architectures for speaker separator networks, Conformer-FFN and TCN-Conformer to perform the target speaker extraction in the time domain. Both architectures were designed to effectively capture both local and global-context information. The first proposed architecture (Conformer-FFN) consists of stacked conformer blocks and external feed-forward blocks, aiming at providing effective feature representation while reducing the overall number of parameters. The second proposed architecture (TCN-Conformer) consists of stacked TCN blocks and conformer blocks, aiming at exploiting the local-context features using TCN blocks first and then exploiting both local and global-context features using conformer blocks. The proposed architectures were evaluated on 2-speaker, 3-speaker, noisy 2-speakers mixtures, comparing their performance against a TCN-based baseline SC-TSE system. Experimental results demonstrated that the proposed TCN-Conformer outperforms both the baseline system and the proposed Conformer-FFN system. The best performance was achieved with TCN-Conformer, utilizing four stacks of TCN and conformer blocks.

Second, we further extended the proposed TCN-Conformer system to make it more suitable for real-time target speaker extraction by replacing the traditional MHSA in each conformer block with linear MHSA. Additionally, a systematic reduction of the overall number of parameters of the system was investigated to further optimize the real-time performance. Experimental results showed that the TCN-Conformer system using linear MHSA consistently outperformed TCN-Conformer systems using traditional MHSA while achieving a substantial reduction in computational cost and RTF.

Future work could explore further optimizations of the proposed systems, such as integrating more efficient MHSA or state-space models for further reducing the computational costs and memory usage or extending this SC-TSE for real-time multi-channel scenarios for spatially-aware target speaker extraction.

5

SUBJECTIVE EVALUATION OF SPEAKER-CONDITIONED TARGET SPEAKER EXTRACTION ALGORITHMS WITH NORMAL-HEARING AND HEARING-IMPAIRED LISTENERS

While Chapters 3 and 4 focus exclusively on objective metrics to evaluate the performance of SC-TSE algorithms, this chapter evaluates the performance of SC-TSE algorithms using subjective evaluation measures based on listening test experiments.

As discussed in Chapter 1, understanding the target speaker in a complex multi-talker scenario requires significantly more cognitive effort compared to a quiet setting. Even NH listeners often struggle to fully understand the target speaker under these conditions [209], [210]. This becomes even more challenging for HI listeners [211], [212] due to peripheral hearing deficits that affect their selective attention [2].

Although objective metrics such as SDR, SI-SDR, PESQ, and STOI have demonstrated the strong performance of these algorithms as observed in Chapters 3 and 4, indicating that SC-TSE algorithms can substantially improve the quality and intelligibility of the target speaker, these metrics do not always fully reflect human perception of speech quality and intelligibility. Therefore, subjective evaluations involving human listeners are essential to assess the effectiveness and benefits of SC-TSE algorithms. A recent study [213] offers valuable insight in this regard by evaluating a quasi-causal SC-TSE algorithm with both NH and HI listeners using double-blind sentence recognition tests. The experiments were conducted in a restaurant noise scenario with mixtures of one, two, or three speakers. The subjec-

This chapter is partly based on the following publication:
[105] R. Sinha, A.-C. Scherer, S. Doclo, C. Rollwage, and J. Rennies, "Subjective performance evaluation of single-channel speaker-conditioned target speaker extraction algorithms for complex acoustic scenes," in *Proc. ITG Conference on Speech Communication*, Aachen, Germany, Sept. 2023, pp. 1–5
[106] R. Sinha, A.-C. Scherer, S. Doclo, C. Rollwage, and J. Rennies, "Evaluation of speaker-conditioned target speaker extraction algorithms for hearing-impaired listeners," *Trends in Hearing*, vol. 29, p. 23 312 165 251 365 802, 2025

tive results of the study [213] demonstrated that both NH and HI listeners benefited from the SC-TSE algorithm, with HI listeners experiencing more benefits compared to NH listeners.

Building on these findings, this chapter focuses on an in-depth subjective evaluation of two fully causal SC-TSE algorithms (referred to as Algo-1 and Algo-2) with both NH and HI listeners, evaluated in unaided (without hearing loss compensation) and aided (with linear hearing loss compensation) conditions. We use three evaluation methods: paired comparison, adaptive measurements of speech recognition thresholds (SRTs), and categorically scaled perceived listening effort across a broader range of SNRs. One goal is to determine whether the considered evaluation methods are suitable for assessing the performance of SC-TSE algorithms under realistic conditions. Additionally, the impact of processing artifacts introduced by SC-TSE algorithms on speech perception for both NH and HI listeners is still unknown. Moreover, it is currently unknown if, or how, hearing loss compensation, typically employed in hearing aids to restore audibility, interacts with these algorithms. Understanding this interaction is essential to determine whether SC-TSE algorithms should be adapted differently for aided and unaided HI listeners.

To address these gaps, this chapter investigates the subjective performance of two SC-TSE algorithms with both NH and HI listeners, focusing on the following research questions:

1. Are the considered subjective evaluation methods suitable for assessing SC-TSE algorithms?
2. What is the performance benefit of SC-TSE algorithms for complex acoustic scenarios with same and different genders of interfering speakers?
3. Do SC-TSE algorithms offer comparable or greater benefits for HI listeners compared to NH listeners?
4. Does hearing loss compensation enhance the benefits of SC-TSE processing for HI listeners, or do listeners without hearing loss compensation experience similar benefits?

To answer these questions, we evaluated the potential of two SC-TSE algorithms (Algo-1 and Algo-2) to enhance the speech perception of the target speaker across a broad range of SNRs. The evaluations covered various acoustic conditions, including scenarios with one or two interfering speakers and with or without gender differences between the target speaker and the interfering speaker(s). Evaluations were conducted under both unaided and aided conditions to assess the effect of hearing loss compensation. Furthermore, we compared the perceptual benefits of these algorithms between NH and HI listeners.

The remainder of this chapter is organized as follows. Section 5.1 revisits the two SC-TSE algorithms (Algo-1 and Algo-2) evaluated in this chapter. Section 5.2 discusses the participants and evaluation stimuli. Section 5.3 discusses the evaluation methods used to assess the performance of SC-TSE algorithms. Section 5.4 presents the results for both NH and HI listeners. Section 5.5 provides answers for the research questions based on evaluation results, and highlights the benefits and limitations. Finally, Section 5.6 provides a summary of this chapter.

5.1 Considered SC-TSE algorithms

In this chapter, we consider two different SC-TSE algorithms: Algo-1 and Algo-2, both algorithms utilize the reference speech of the target speaker as auxiliary information. In this section, we briefly discuss these algorithms.

Algo-1
Algo-1 utilizes a ResNet-GRU-based architecture (see Section 2.3) to perform target speaker extraction in the time-frequency domain. The speaker embedder network in Algo-1 utilizes the same architecture as in Chapter 3 to generate 256-dimensional target speaker embedding, while the speaker separator network utilizes a ResNet-GRU architecture to estimate a real-valued mask. The ResNet-GRU speaker separator network consists of two ResNet layers, two uni-directional GRU layers, and two fully connected (FC) layers. Each ResNet layer consists of two basic blocks, with each block containing two CNN layers followed by batch normalization and ReLU activation. The number of nodes in the GRU layers and the first FC layer is fixed to 256, while the last FC layer consists of 257 nodes and uses a sigmoid activation function. For Algo-1, the total computational complexity in terms of number of MACS is 4.1 G, with an algorithmic latency of 32 ms.

Algo-2
Algo-2 utilizes the TCN-Conformer architecture with traditional MHSA, as proposed in Section 4.2.2.3 to perform target speaker extraction in the time domain. The speaker embedder network in Algo-2 utilizes the same ResNet-based architecture as in Chapter 4 to generate 256-dimensional target speaker embedding. The speaker separator network also follows the TCN-Conformer architecture described in Section 4.2.2.3, with two key modifications: first, causal masking is applied to each traditional MHSA at each time step in each conformer block, and second, the padding in the 1D CNN layers is changed from same to causal. All other (hyper)parameters are kept the same as discussed in Section 4.3.2. The input and convolutional size of each TCN block is fixed to 512, and the kernel size is fixed to 3. Each conformer block utilizes 8-head attention with a convolutional kernel size of 31. The output size of the feed-forward layer in each conformer block is 4 times the input size. The feed-forward layer is followed by a Swish activation and a dropout layer. For Algo-2, the total computational complexity in terms of number of MACS is 16.9 G, with an algorithmic latency of 2.5 ms.

Both algorithms were trained on the same dataset as discussed in Section 4.3.1. Algo-1 was trained using the SI-SDR loss function, while Algo-2 was trained on the weighted combination of multi-scale SI-SDR loss and the CE loss functions as discussed in Section 4.3.2. It should be noted that both algorithms were trained and validated on datasets, using mixtures composed of English speech, while the subjective evaluations were conducted using German speech materials.

5.2 Participants and stimuli

5.2.1 *Participants*

NH listeners:
Fifteen native German-speaking NH listeners (6 males, 9 females), aged 19 to 32 years, participated in the listening test experiments. Each participant had normal hearing according to clinical audiometry.

HI listeners:
Fifteen native German-speaking HI listeners (7 males, 8 females), aged 54 to 65 years, participated in the listening test experiments. Fig. 5.1 shows the group mean and individual audiograms for the right and left ears. All participants underwent laboratory-based audiometric testing to confirm mild to moderate sensorineural hearing loss [214] based on their pure-tone thresholds. The hearing loss was relatively symmetric (the difference in pure-tone average between the left and right ears was less than 10 dB for all participants, except one, who showed a difference of 20 dB).

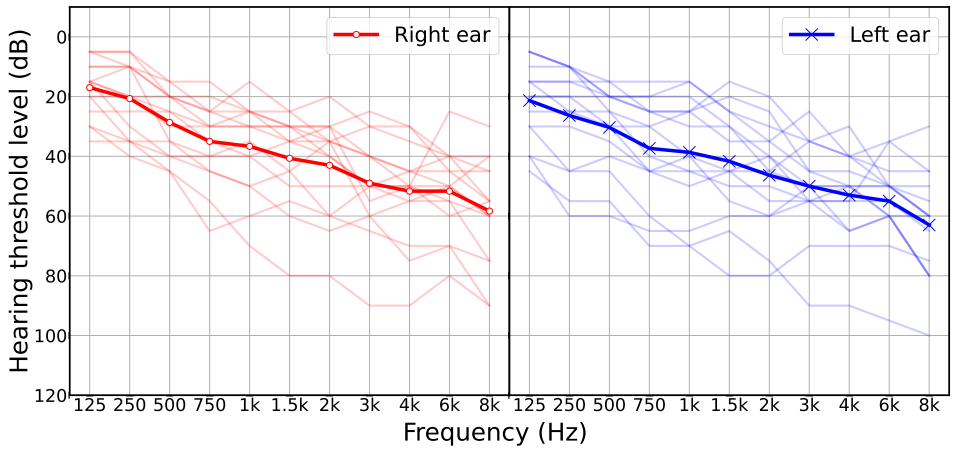

Fig. 5.1: Individual and group mean hearing thresholds (in dB) for the right and left ears for the HI listeners.

All participants from both groups (NH and HI listeners) received hourly compensation and gave informed consent for their participation in the experiments. The methods were approved by the ethics committee of the University of Oldenburg (protocol Drs.EK/2019/073-02).

5.2.2 *Stimuli and equipment*

We used the same evaluation stimuli for both NH and HI listeners. The target speaker stimuli consisted of German matrix sentences uttered by a fixed male speaker from the Oldenburg sentence test (OLSA) [189]. The reference speech of the

target speaker was chosen from the German Göttingen sentence test (GÖSA) [215] which consists of everyday sentences uttered by the same male speaker. Each OLSA sentence followed a fixed syntactical structure containing 5 words in the following order: name, verb, numeral, adjective, and object. For each word, 10 alternatives were available, which were randomly combined to generate syntactically correct but semantically unpredictable sentences. The interfering speech also consisted of matrix sentences uttered by one or two different speakers (either male or female), chosen from the dataset used in [216]. Interfering speaker signals were generated by concatenating several sentences starting at a random position for each presentation. The relative level of the target speaker and the interfering speaker(s) was varied to produce different SNRs. As the reference speech of the target speaker, several utterances of the target speaker from the GÖSA sentence test were concatenated to make a 10s-long utterance. Although both SC-TSE algorithms were also trained with noisy mixtures of two speakers, no such stimuli were included in the evaluation experiments.

To familiarize the participants with the voice of the target speaker, each participant listened to an example of about $60s$ consisting of concatenated sentences uttered by the target speaker. These sentences were mixed with interfering speakers as in the experiments, but at a high SNR (between +5 to +10 dB) to ensure that the target speaker was much louder than the interfering speakers. During the experiments, stimuli were presented diotically via Sennheiser HD650 headphones in sound-attenuated booths.

5.3 Subjective evaluation methods and procedures

In this section, before presenting the procedure followed for each evaluation method in detail, we first provide a brief discussion on how the stimuli were adjusted for both groups of listeners.

To assess the performance of both algorithms (Algo-1 and Algo-2) with both NH listeners and HI listeners, paired comparisons [188], speech intelligibility measurements [189], and perceived listening effort scaling [190] were utilized. These methods vary in terms of the outcome measure and the SNR range to which they are applicable. Paired comparisons were used to determine the preferences of participants between different versions of the same stimulus. An SNR=0 dB was used because this test scenario is typically considered in objective evaluations of SC-TSE algorithms. Speech intelligibility was measured in terms of SRTs, i.e., SNRs corresponding to 50% speech intelligibility. SNRs are typically negative at such low-performance levels, at least for NH listeners (as observed in, e.g., [210], [217]). Categorical listening effort scaling was used to assess the perceived effort that a participant needed to understand the target speaker. This method is typically measured over a broad range of SNRs.

For both NH and HI listeners, paired comparisons and perceived listening effort were measured for stimuli in which the target speaker was masked by either one or two interfering speakers, while SRTs were measured only for stimuli having two interfering speakers. We excluded SRT measurements with only one interfering speaker,

as SRTs for such conditions are known to be extremely low [218], i.e., falling into SNR regions where algorithms are not expected to perform well, nor where typical listening conditions would occur [219]. Evaluation with each method was conducted for three processing conditions (Unprocessed, Algo-1, Algo-2), two genders of interfering speakers (male and female), and different numbers of interfering speakers (depending on the evaluation method). In the paired comparison, all combinations of three pairwise processing conditions, one and two interfering speakers, and two genders were considered, resulting in 12 unique conditions. Each condition was repeated three times, leading to 36 trials per participant. For the SRT measurements, only two interfering speakers were considered, combined with three processing conditions and two genders, resulting in 6 unique conditions and approximately 120 trials per participant. The listening effort measurement included all combinations of three processing conditions, one and two interfering speakers, two genders, and five SNRs, resulting in 60 unique conditions. Each condition was repeated three times, leading to 180 trials per participant.

Furthermore, for evaluation with NH listeners, the signals were scaled such that the target speaker had the same level (70 dB SPL) as the single or the combined interfering speaker(s). For evaluation with HI listeners, the signals were initially scaled such that the target speaker had the same level (65 dB SPL before hearing loss compensation) as the single or the combined interfering speaker(s), and then the target speaker was adapted to generate stimuli at different SNRs across all evaluation methods. In the processed conditions, these mixtures were processed by either Algo-1 or Algo-2, typically reducing the level of the interfering speaker(s) energy. In the aided conditions for HI listeners, hearing loss compensation was applied after processing.

5.3.1 *Paired comparisons*

Fig. 5.2: Visual representation of paired comparison test panel.

In each trial, two versions of the same stimulus were presented to the participants, visually labeled as intervals A and B. The participants could toggle between these two versions as many times as they liked by clicking on the intervals. The stimuli were played in a loop, allowing participants to decide in which interval the target

speaker was more intelligible. Participants were asked to rate on a six-point scale (see Fig. 5.2) if one interval was "much easier" (German: "viel einfacher") (A+++ / B+++), "clearly easier" ("deutlich einfacher") (A++ / B++), or "easier" ("einfacher") (A+ / B+) intelligible than the other. The middle category ($A = B$) on the rating scale was deliberately omitted, requiring participants to select one interval in each trial. All three versions (unprocessed, stimuli processed by Algo-1, and stimuli processed by Algo-2) were compared with each other, while in each trial the assignment to intervals A and B was randomized. For each comparison, three repetitions were performed using different target and interfering sentences. The outcomes of each comparison were analyzed in terms of the percentage of wins.

5.3.2 *Speech recognition thresholds*

Speech recognition thresholds (SRTs) were measured using an adaptive procedure. In each trial, participants were presented with a mixture of the target speaker and two interfering speakers (either processed or unprocessed) once. They were then asked to select the recognized words from a word matrix displayed on the screen before proceeding to the next trial (see Fig. 5.3).

Fig. 5.3: Visual representation of speech recognition thresholds measurement test.

For NH listeners, the level of the combined interfering speakers was fixed at 70 dB SPL, while the level of the target speaker was adjusted adaptively based on the participant responses in the previous trial. For HI listeners, the level of the combined interfering speakers was fixed at 65 dB SPL (before hearing loss compensation),

while the level of the target speaker was adjusted adaptively based on the participant responses in the preceding trial. The initial SNR was set to 5 dB, and the step size was varied according to the adaptive procedure proposed in [220] to converge to the SRT. To prevent clipping and excessively loud stimuli, the maximum SNR was limited to 20 dB.

For both NH and HI listeners, If a participant correctly identified three or more words out of five, the SNR was decreased, otherwise, the SNR was increased. Each version of the stimuli (unprocessed, processed by Algo-1, and processed by Algo-2) was evaluated using a different list of 20 distinct sentences, presented in random order. To minimize training effects, participants completed two training lists before the actual SRT measurements. These lists contained 20 sentences spoken by the target speaker, mixed with stationary noise, following the same procedure as in [189].

5.3.3 *Perceived listening effort*

In each trial, a mixture of the target speaker and interfering speaker(s) (processed or unprocessed) was presented to the participants. The participants were asked to rate the perceived effort required to understand the target speaker on a 13-point scale ranging from "no effort" (German: "mühelos") corresponding to 1 effort scaling categorical unit (ESCU) to "extreme effort" ("extrem anstrengend") (13 ESCU). An additional fourteenth category "only interfering speakers" ("nur Störsprecher") was included (see Fig. 5.4) for trials in which participants could only hear the interfering speaker(s).

These assessments employed stimuli with predetermined SNRs as in [190]. During each trial, the stimulus was played continuously in a loop until participants provided their rating, after which the next trial started. All SNRs and processing conditions were presented in a random order. The target and interfering speaker(s) were mixed at SNRs ranging from −10 dB to 15 dB, with a step size of 5 dB. The overall stimulus level was kept fixed at 70 dB SPL for NH listeners to avoid large differences in loudness between trials, while it was kept fixed at 65 dB SPL (before hearing loss compensation) for HI listeners. For each combination of SNR, processing condition, and interfering speaker condition (one or two), the measurement was repeated three times using distinct sentences. The median value obtained from these repetitions was utilized as the assessment of an individual's perceived listening effort for that specific combination.

All experiments were performed with both NH listeners and HI listeners, where both unaided and aided conditions were considered separately for evaluation with HI listeners. For the unaided condition, no hearing loss compensation was provided, while for the aided condition, individualized amplification was provided to the participants. The stimuli were amplified by applying a linear gain according to the National Laboratories Revised Profound (NAL-RP) prescription [221]. For each participant, the amplification applied to the left and right ears was identical and calculated based on the average hearing threshold across both ears.

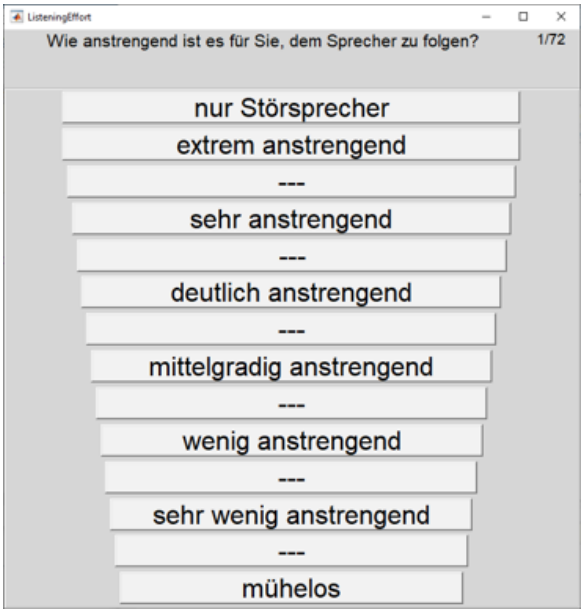

Fig. 5.4: Visual representation of perceived listening effort measurement scale.

5.4 Evaluation results

In this section, we compare the performance of Algo-1 and Algo-2 against the unprocessed mixture based on paired comparisons, SRTs, and perceived listening effort, with both NH and HI listeners. Section 5.4.1 presents the evaluation results for NH listeners and Section 5.4.2 presents the evaluation results for HI listeners, considering both unaided and aided conditions.

5.4.1 *Subjective evaluation results with NH listeners*

5.4.1.1 *Paired comparisons with NH listeners*

Fig. 5.5 shows the percentage of wins from the paired comparison tests between the unprocessed stimuli and the stimuli processed by each algorithm. The top and middle panels compare the unprocessed stimuli with the stimuli processed by Algo-1 and Algo-2, while the bottom panel compares Algo-1 directly with Algo-2. Different hatches/colors represent different masking conditions (M/F: one male/female interfering speaker, MM/FF: two male/female interfering speakers). The data reveal a very clear preference for Algo-2 compared to the unprocessed stimuli: 100% of all comparisons favor processed stimuli for a single interfering speaker. For two interfering speakers, the percentage of wins was about 96% (F) and 98% (M). The distribution of ratings indicates that in most cases processed stimuli using Algo-2 were perceived as "clearly easier" (++) to understand than the unprocessed stim-

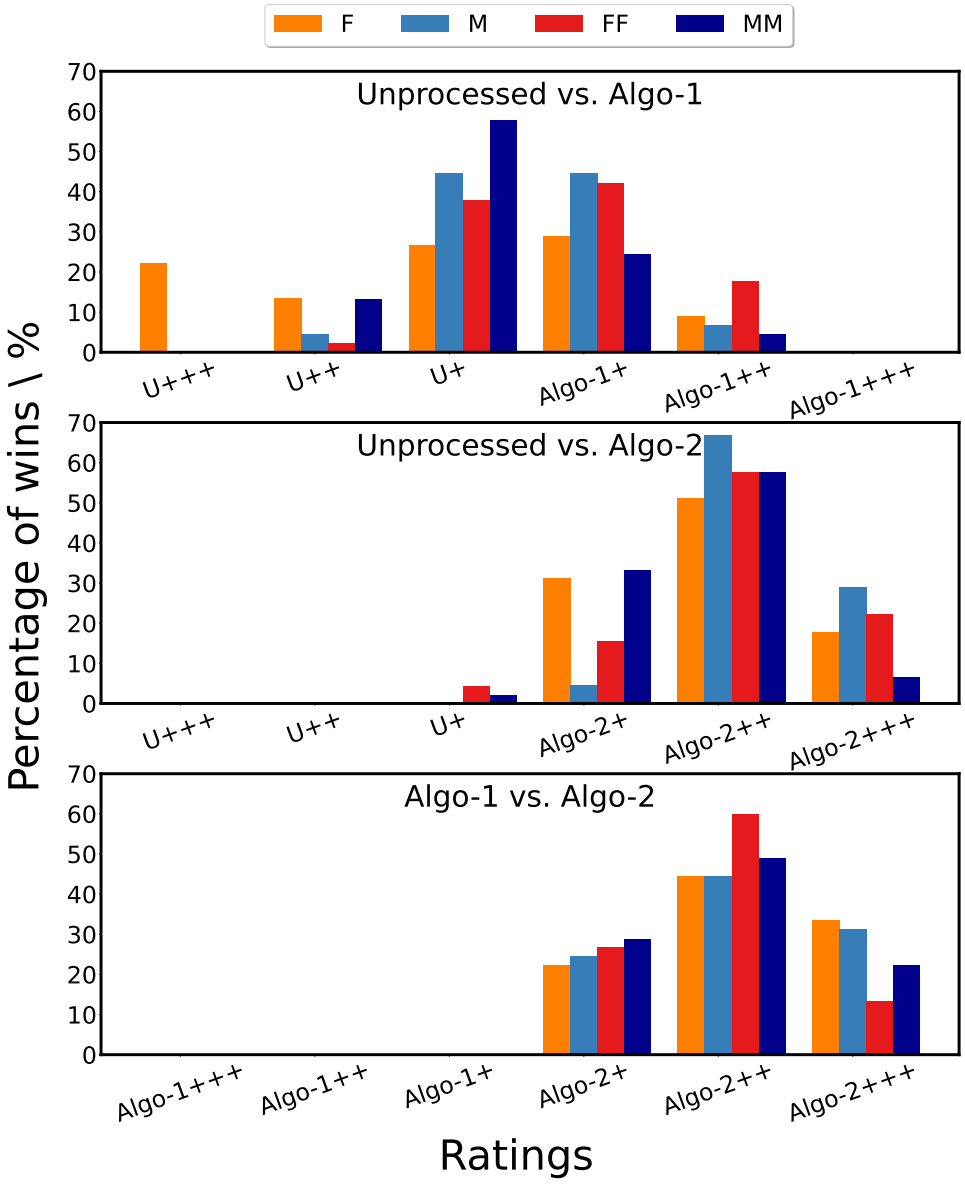

Fig. 5.5: Percentage of wins from the paired comparison tests obtained for each pair of the three processing conditions (unprocessed, Algo-1, and Algo-2) for stimuli having one (F/M) or two (FF/MM) interfering speakers with NH listeners.

uli. In contrast, Algo-1 did not provide a consistent advantage over the unprocessed stimuli. Most ratings were given to the middle categories of the rating scale, indicating that listeners were uncertain whether the unprocessed or the processed stimuli were easier to understand. Consistently, Algo-2 was strongly preferred over Algo-1: not a single rating favored Algo-1, and most ratings indicated that stimuli processed by Algo-2 were "clearly easier" to understand.

5.4.1.2 *Speech recognition thresholds with NH listeners*

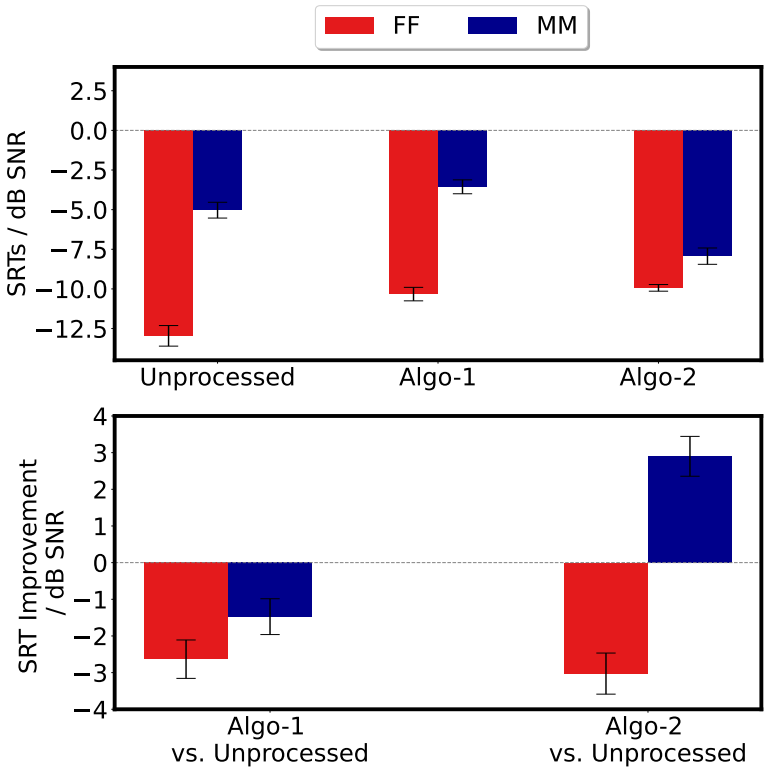

Fig. 5.6: SRTs averaged across all participants (top) and corresponding SRT improvements (bottom) obtained for stimuli having two female(FF) or male (MM) interfering speakers with NH listeners. Error bars represent standard errors.

Fig. 5.6 shows measured SRTs averaged across participants for the three processing (unprocessed, Algo-1, and Algo-2), where error bars represent standard errors. In general, SRTs were negative for all processing (indicating that listeners were able to understand 50% of the target speaker when the energy of the interfering speakers exceeded the energy of the target speaker) and significantly lower for female than male interfering speakers. For male interfering speakers, the highest mean SRT achieved by Algo-1 is −3.6 dB, while Algo-2 improves SRTs (−7.9 dB) compared to the unprocessed stimuli (−5.0 dB). For female interfering speakers, the lowest

SRT for the unprocessed stimuli was −13 dB, while SRTs for the processed stimuli were approximately 3 dB higher for both processed conditions. Additionally, the statistical significance was tested using a two-way analysis of variance for repeated measures, followed by Bonferroni-corrected t-tests as post-hoc tests. The factor processing ($F(2, 28) = 18.833, p<0.001$) as well as the interaction between processing and the gender of interfering speakers ($F(2, 28) = 41.003, p<0.001$) also significantly affected the performance. All pairwise tested differences were significant ($p \leq 0.002$) at the corrected significance level (0.05/9) except for unprocessed stimuli with male interfering speakers and corresponding processed stimuli with Algo-1 and processed stimuli having female interfering speakers using Algo-1 vs. Algo-2.

5.4.1.3 *Perceived listening effort with NH listeners*

Fig. 5.7 shows median listening effort ratings across participants along with the corresponding benefits for one and two interfering speakers as a function of SNR. The first three rows represent the listening effort ratings for the unprocessed stimuli, Algo-1, and Algo-2, while the last row represents the listening effort benefits of both algorithms compared to the unprocessed stimuli. For the unprocessed stimuli, perceived listening effort decreased systematically with increasing SNR. At the two lowest SNRs, listening effort was higher for two interfering than for one interfering speaker. For Algo-1, perceived effort was similar to the unprocessed stimuli for lower SNRs, and considerably higher at high SNRs. In contrast, Algo-2 showed a consistent decrease in listening effort compared to the unprocessed stimuli at all SNRs for both genders of interfering speakers except at −10 dB. The mean benefit for Algo-1 and Algo-2 compared to the unprocessed stimuli (see bottom panel of Fig. 5.7) was the largest at intermediate SNRs and reached more than 4 categories on the 13-point scale for a single interfering speaker. For two interfering speakers, the benefit had a similar pattern but was slightly lower than a single interfering speaker. This was also confirmed by single-sample t-tests conducted to test if the mean benefits differed significantly from 0. For Algo-2, there was a significant reduction in listening effort for SNRs from −5 to 10 dB for both male and female interfering speakers. For two interfering speakers, Algo-1 significantly increased listening effort at +15 dB for male interfering speakers and 5 dB for female interfering speakers, while Algo-2 significantly improved listening effort at 0 and 10 dB for male interfering speakers, and at −5, 0, and 10 dB for female interfering speakers.

5.4.2 *Subjective evaluation results with HI listeners*

5.4.2.1 *Paired comparisons with HI listeners*

Fig. 5.8 shows the percentage of wins from the paired comparison tests for both unaided (left column) and aided (right column) conditions. The top and middle panels compare the unprocessed stimuli with the stimuli processed by Algo-1 and Algo-2, while the bottom panels compare Algo-1 directly with Algo-2. Different hatches/colors represent different masking conditions (M/F: one male/female interfering speaker, MM/FF: two male/female interfering speakers). For both unaided

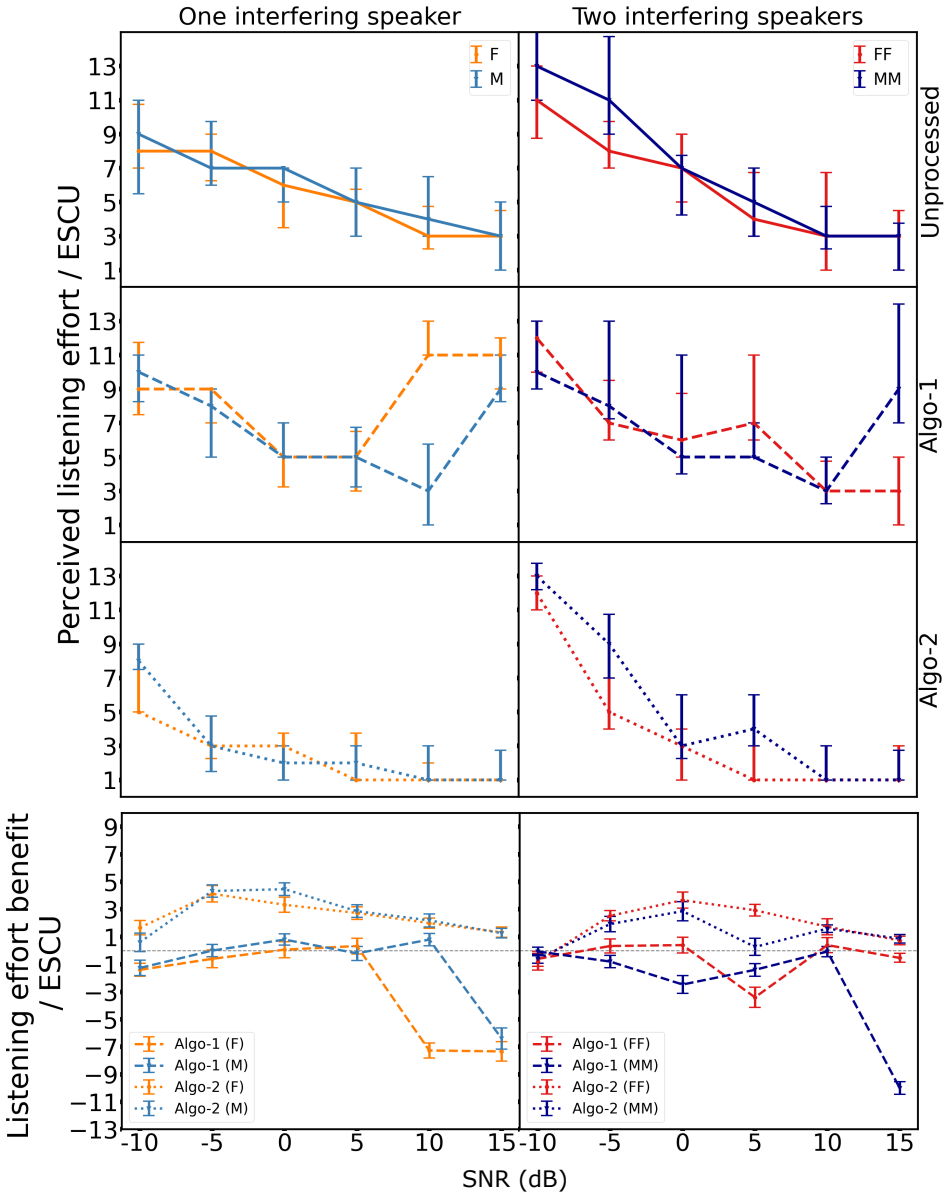

Fig. 5.7: Median perceived listening effort ratings and benefit relative to unprocessed stimuli as a function of SNR for stimuli having one (F/M) or two (FF/MM) interfering speaker(s) with NH listeners. The first three rows represent the listening effort ratings for unprocessed stimuli, Algo-1, and Algo-2, while the last row represents the listening effort benefit of Algo-1 and Algo-2 compared to unprocessed stimuli. Error bars represent interquartile difference for the perceived listening effort ratings, and standard errors for the listening effort benefit.

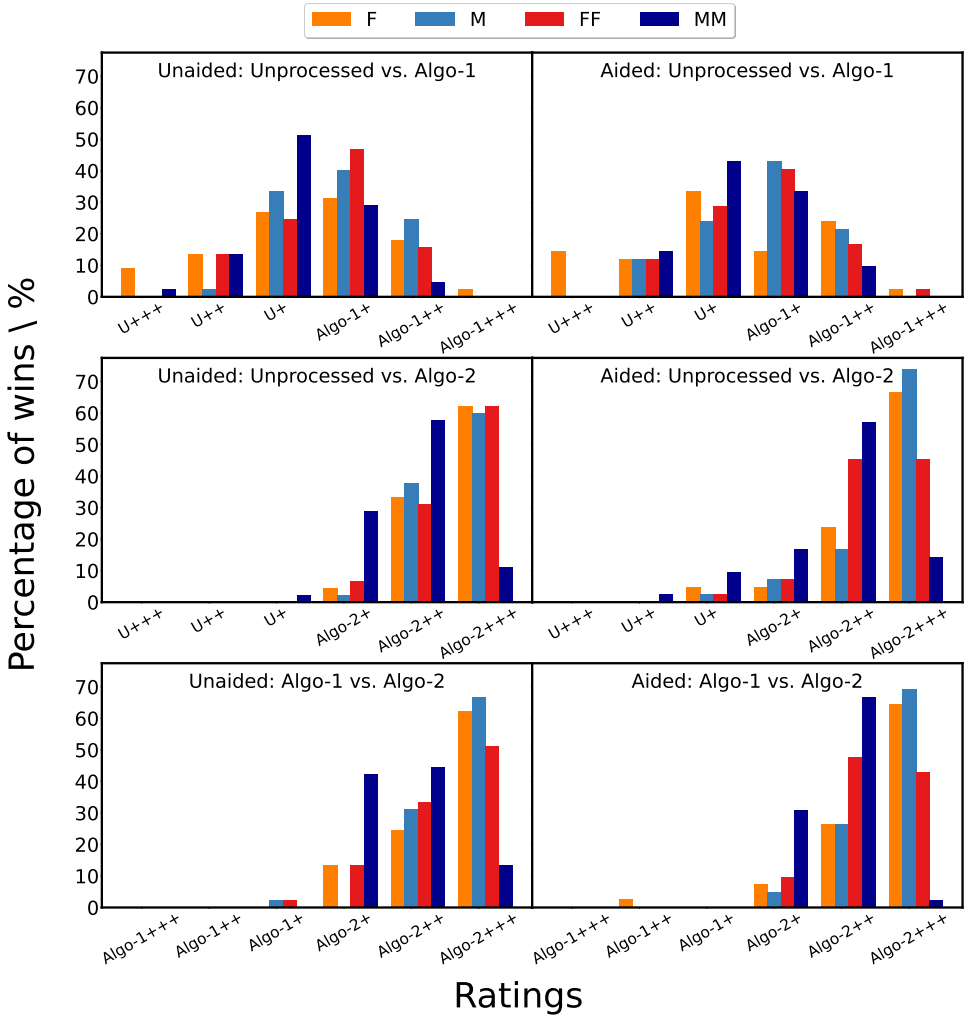

Fig. 5.8: Percentage of wins from the paired comparison tests obtained for each pair of the three processing conditions (unprocessed, Algo-1, and Algo-2) for stimuli having one (F/M) or two (FF/MM) interfering speakers. The left column represents the ratings for the unaided conditions with HI listeners, and the right column the ratings for the aided conditions with HI listeners.

and aided conditions, the data reveal a relatively similar pattern of ratings, where a clear preference for Algo-2 was observed compared to the unprocessed stimuli and Algo-1. Stimuli processed by Algo-2 were favored in comparison to the unprocessed stimuli in 100% of all comparisons, and in 98% for two male (MM) interfering speakers. The category "much easier" (+++) was given most often for one male/female and two female interfering speaker(s), while "clearly easier" (++) was given most often for two male interfering speakers. Similarly, in the direct comparison between the algorithms, participants frequently rated Algo-2 as "much easier" to understand than Algo-1. In neither unaided nor aided conditions, participants found any benefit of Algo-1 compared to the unprocessed stimuli as most ratings were given to the middle categories of the rating scale, indicating that participants were uncertain about making a decision.

5.4.2.2 *Speech recognition thresholds with HI listeners*

Fig. 5.9 shows the measured averaged SRTs (top panels) and the corresponding improvements achieved by each algorithm compared to the unprocessed stimuli (bottom panels). For both unaided (left column) and aided (right column) conditions, mean SRTs were considerably lower for female interfering speakers than for male interfering speakers. Mean SRTs obtained for the unprocessed stimuli were −5.1 dB (female) and 0.0 dB (male) for the unaided condition, and −5.9 dB (female) and −1.3 dB (male) for the aided condition. For both unaided and aided conditions, Algo-1 showed no benefits in mean SRTs (even an increase) compared to the unprocessed stimuli, whereas Algo-2 showed considerable benefits for both female and male interfering speakers. For the unaided condition, Algo-2 achieved 1.2 dB (female) and 3.2 dB (male) lower SRTs. For the aided condition, the benefit was 1.6 dB (female) and 2.9 dB (male). For the aided condition, mean SRTs were negative for all three processing conditions (unprocessed, Algo-1, and Algo-2), indicating that participants were able to understand 50% of the target speaker even when the energy of the interfering speakers exceeded that of the target speaker.

It should be noted that, for the unaided condition, the SRT measurement of three participants were invalid for Algo-1 with male interfering speakers. This occurred due to participants mistakenly followed one of the interfering speakers instead of the target speaker. As a result, the adaptive procedure kept increasing the SNR until the predefined maximum of 20 dB was reached. At three ceiling hits, the trial was aborted automatically.

Statistical analyses were performed using a linear mixed-effects model with the lme4 package in R software [222], which is well-suited for handling missing data (invalid data from the three participants were considered as missing data). Participants were treated as a random factor. We conducted a comprehensive diagnostic evaluation, including visual inspection of posterior predictions, linearity, homogeneity of variance, collinearity, influential observations, normality of residuals, and normality of random effects using the performance package in R [223]. Furthermore, we performed contrast analysis with Holm corrections using the model-based package [224], with an alpha level of 0.05 for all tests. Visual inspections of the residuals

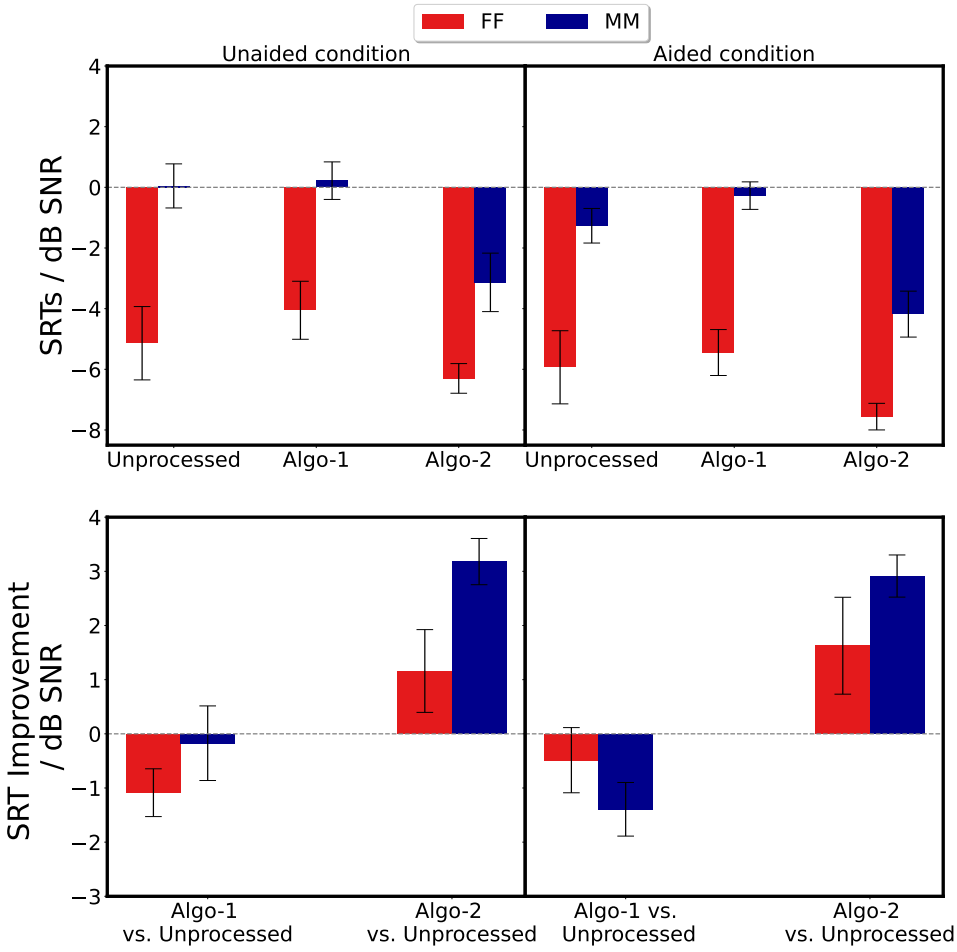

Fig. 5.9: SRTs averaged across all participants (top) and corresponding SRT improvements (bottom) obtained for stimuli having two female(FF) or male (MM) interfering speakers. The left column represents SRTs and corresponding improvements for the unaided conditions with HI listeners, and the right column the aided conditions with HI listeners. Error bars represent standard errors.

of the linear mixed-effects model predicting the outcome of variable SRTs revealed normal distribution.

A linear mixed-effects model was fitted to the measured SRTs to estimate the fixed effects of processing (unprocessed, Algo-1, and Algo-2), hearing loss compensation (unaided and aided), and the gender of interfering speakers (female and male), along with their two- and three-way interactions. An analysis of variance revealed significant main effects of processing ($F(2, 154) = 44.6, p < 0.001$), hearing loss compensation ($F(1, 154) = 16.2, p < 0.001$), and the gender of interfering speakers ($F(1, 154) = 267.7, p < 0.001$). Additionally, a significant two-way

Table 5.1: Results of contrast analysis for predicting the differences in SRTs. Only significant differences are reported.

Part-1

Effect	**Difference (dB SNR)**	**p-Value**
Male | Female	4.3	$< .001$
Unaided | Aided	1.1	$< .001$
Algo-1 | Algo-2	2.9	$< .001$
Unprocessed | Algo-1	-0.7	.030
Unprocessed | Algo-2	2.2	$< .001$

Part-2

Effect	**Interfering Speakers**	**Difference (dB SNR)**	**p-Value**
Algo-1 | Algo-2	Female	2.2	$< .001$
Algo-1 | Algo-2	Male	3.7	$< .001$
Unprocessed | Algo-2	Female	1.4	.005
Unprocessed | Algo-2	Male	3.1	$< .001$
Effect	**Processing**	**Difference (dB SNR)**	**p-Value**
Male | Female	Algo-1	4.8	$< .001$
Male | Female	Algo-2	3.3	$< .001$
Male | Female	Unprocessed	4.9	$< .001$

interaction was found between processing and the gender of interfering speakers ($F(2, 154) = 3.9, p = 0.021$). However, neither a statistically significant three-way interaction nor an interaction between hearing loss compensation and processing were found.

Significant effects were further analyzed using contrast analysis to compare the three factors (see Table 5.1). Table 5.1 is divided into two parts. The first part presents the main effects of processing, hearing loss compensation, and the gender of interfering speakers. The second part presents the pairwise differences between all levels of processing for each gender of interfering speakers and the pairwise differences between all levels of gender for each processing. Only statistically significant differences are reported. The results revealed significant differences between unprocessed and Algo-2, as well as between Algo-1 and Algo-2, for both male and female interfering speakers. However, no significant difference was found between unprocessed

We also explored using mean, median, and multiple data imputation approaches rather than treating the data as missing in the linear mixed-effects models to assess whether different handling of missing data would affect the analysis. The results remained consistent.

and Algo-1. Additionally, for each processing (unprocessed, Algo-1, and Algo-2), a significant difference was observed between male and female interfering speakers.

5.4.2.3 *Perceived listening effort with HI listeners*

Fig. 5.10 shows the median listening effort ratings across participants along with the corresponding benefits for one and two interfering speakers for the unaided condition, while Fig. 5.11 shows the median listening effort ratings across participants along with the corresponding benefits for one and two interfering speakers for the aided condition as a function of SNR.

The first three rows represent the listening effort ratings for the unprocessed stimuli, Algo-1, and Algo-2, while the last row represents the listening effort benefits of both algorithms compared to the unprocessed stimuli in each Fig. 5.10 and 5.11. In general, listening effort ratings systematically decreased with increasing SNR for both one and two interfering speakers (except for Algo-1 at high SNRs), and followed a similar pattern for unaided and aided conditions. For the unprocessed stimuli and low SNRs, the perceived effort was higher with two interfering speakers compared to one interfering speaker. For two interfering speakers, participants also rated the 14-th category "only interfering speakers" for male (unaided) and for female (aided) at the lowest SNR (−15 dB). Algo-1 showed a minimal reduction in listening effort at 5 dB SNR (one interfering speaker), but no reduction at other SNRs. At higher SNRs, it even increased the listening effort compared to the unprocessed stimuli, likely due to artifacts such as residuals of interfering speakers that affect the overall quality and intelligibility of the processed signal. In contrast, Algo-2 reduced perceived listening effort at all considered SNRs for both one and two interfering speakers. Participants gave a median rating of "no effort" for one interfering speaker (male and female) at every SNR except at −10 dB, and for two female interfering speakers except at −10 and −5 dB. Overall, Algo-2 showed significant benefits at all considered SNRs compared to the unprocessed stimuli for both one and two interfering speakers.

To investigate the effect of SNR and hearing loss compensation on listening effort ratings, we conducted a statistical analysis of the listening effort benefit provided by Algo-1 and Algo-2 compared to the unprocessed stimuli. A linear mixed-effects model was fitted to the listening effort benefit to estimate the fixed effects of processing benefits (Algo-1 vs. unprocessed, and Algo-2 vs. unprocessed), hearing loss compensation (unaided and aided), and SNRs, along with their two-way interactions. An analysis of variance revealed significant main effects for processing benefits ($F(1, 1412) = 789.8, p < 0.001$) and SNRs ($F(5, 1412) = 118.2, p < 0.001$), as well as a significant two-way interaction between processing benefits and SNRs ($F(5, 1412) = 23.1, p < 0.001$). However, no statistically significant main effect or two-way interactions including hearing loss compensation were observed. The significant effects and interactions were further analyzed using contrast analysis (see Table 5.2).

Table 5.2 is divided into two parts. Part 1 presents the main effect of processing benefit, while Part 2 presents the pairwise comparison of processing benefit for each SNR. Only statistically significant differences are reported. The main effects of SNR

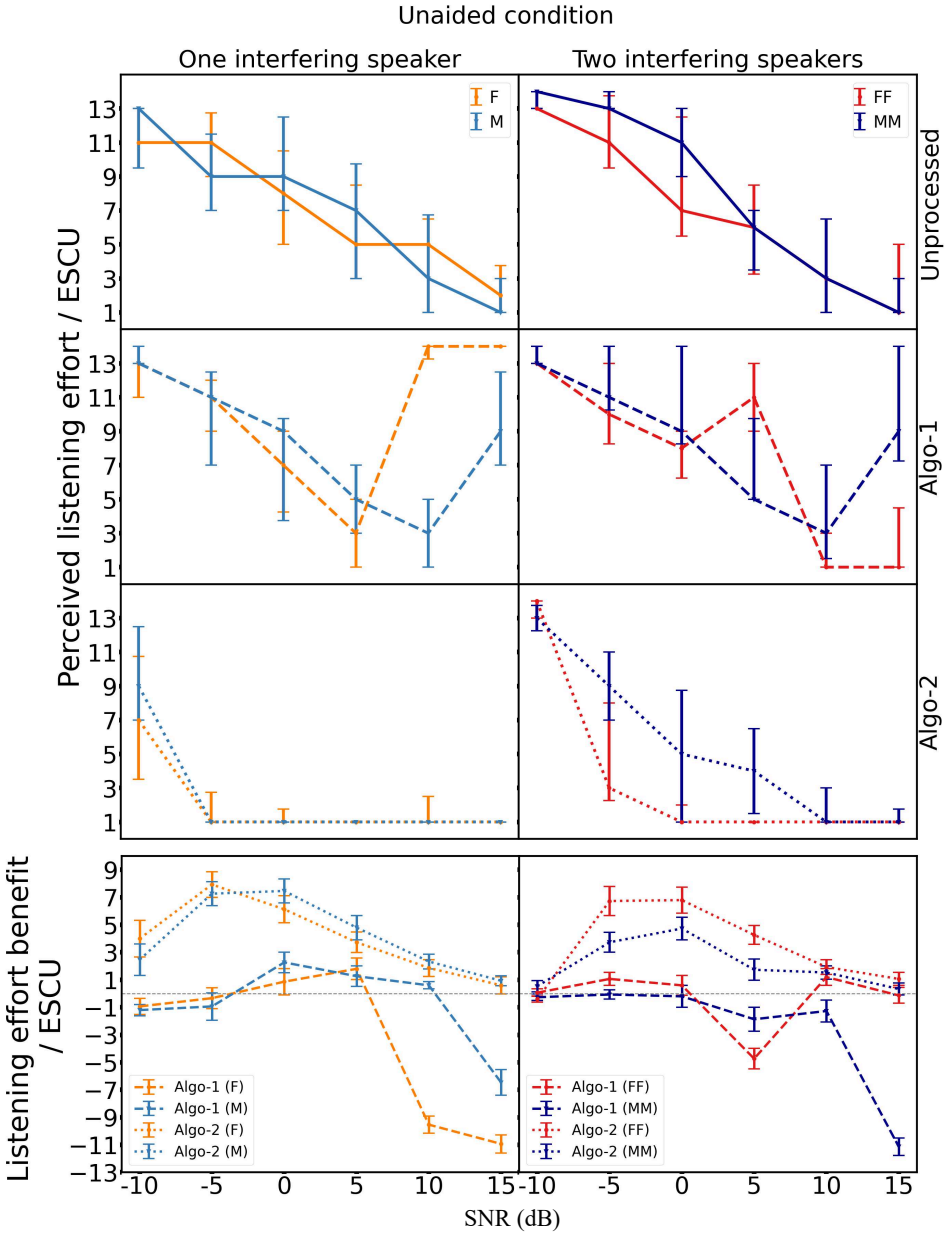

Fig. 5.10: Median perceived listening effort ratings and benefit relative to unprocessed stimuli as a function of SNR for stimuli having one (F/M) or two (FF/MM) interfering speaker(s) for the unaided condition with HI listeners. The first three rows represent the listening effort ratings for unprocessed stimuli, Algo-1, and Algo-2, while the last row represents the listening effort benefit of Algo-1 and Algo-2 compared to unprocessed stimuli. Error bars represent interquartile difference for the perceived listening effort ratings, and standard errors for the listening effort benefit.

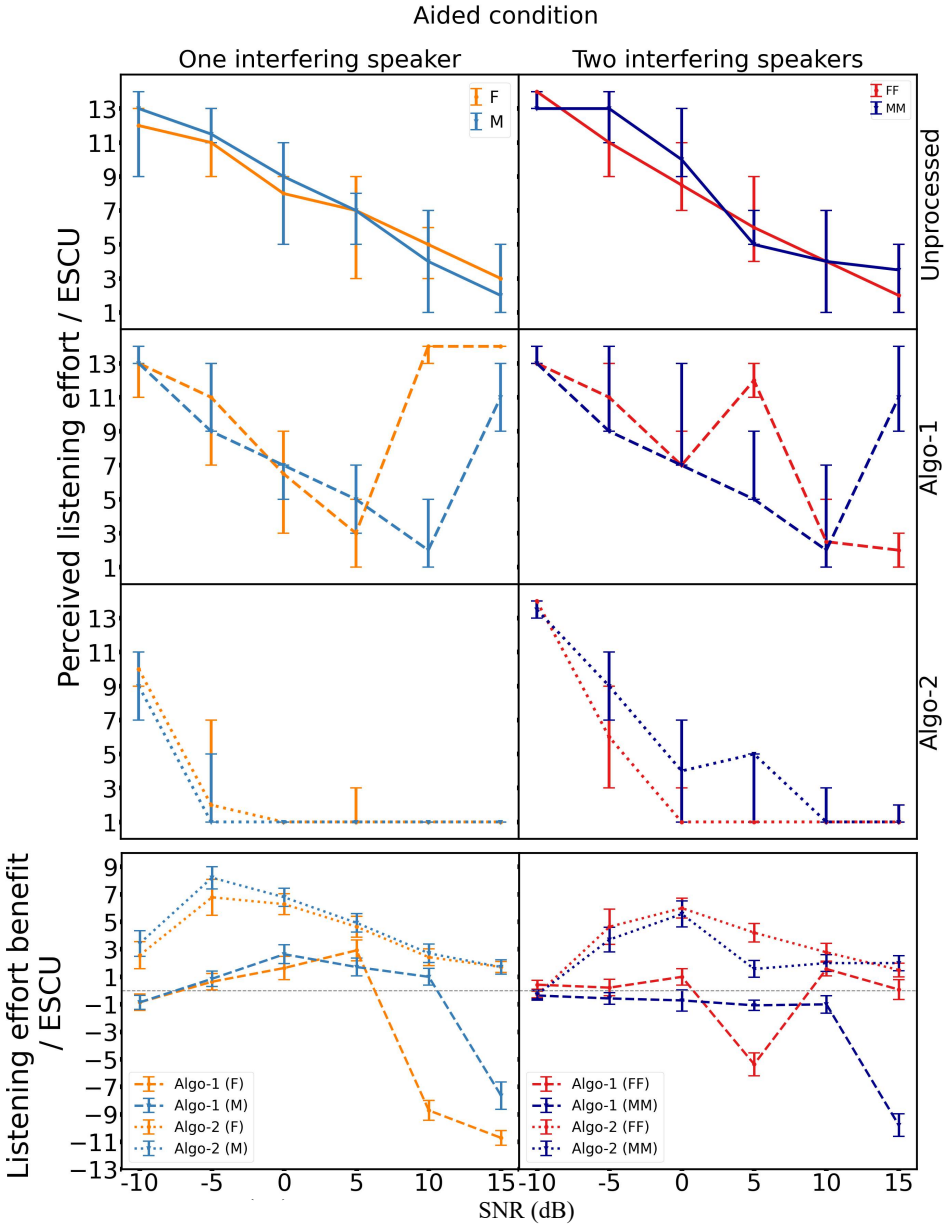

Fig. 5.11: Median perceived listening effort ratings and benefit relative to unprocessed stimuli as a function of SNR for stimuli having one (F/M) or two (FF/MM) interfering speaker(s) for the aided condition with HI listeners. The first three rows represent the listening effort ratings for unprocessed stimuli, Algo-1, and Algo-2, while the last row represents the listening effort benefit of Algo-1 and Algo-2 compared to unprocessed stimuli. Error bars represent interquartile difference for the perceived listening effort ratings, and standard errors for the listening effort benefit.

and the pairwise comparisons between all SNRs for each processing benefit type were also analyzed. The results revealed that, for each SNR, there was a significant difference in the benefits provided by Algo-1 and Algo-2. Additionally, significant differences were found between most SNR pairs, except for the following cases: (10 dB vs. −10 dB, 15 dB) for Algo-2; (−5 dB vs. −10 dB, 5 dB) for Algo-1; (−5 dB vs. 0 dB) for both Algo-1 and Algo-2; (−10 dB vs. 10 dB) for Algo-2; and (−10 dB vs. −5 dB) for Algo-1.

Table 5.2: Results of contrast analysis for predicting the differences in listening effort benefits. Only significant differences are reported.

Part-1

Effect	**Benefit difference**	**p-Value**
Algo-1 | Algo-2	-5.0	$< .001$

Part-2

Effect	**SNR (dB)**	**Benefit difference**	**p-Value**
Algo-1 | Algo-2	-10	-2.0	$< .001$
Algo-1 | Algo-2	-5	-6.0	$< .001$
Algo-1 | Algo-2	0	-5.2	$< .001$
Algo-1 | Algo-2	5	-4.4	$< .001$
Algo-1 | Algo-2	10	-4.2	$< .001$
Algo-1 | Algo-2	15	-8.3	$< .001$

5.5 Discussion

5.5.1 *Differences for unprocessed stimuli between NH and HI listeners*

The data revealed considerable differences in terms of SRTs and listening effort ratings between NH (see Figs. 5.6 and 5.7) and HI listeners (see Figs. 5.9, 5.10, and 5.11) for unprocessed stimuli. SRTs for HI listeners were considerably higher than SRTs of NH listeners with a difference of 5 dB for male interfering speakers and 8 dB for female interfering speakers. Similar findings for SRT measurements with HI listeners were also reported in [225] in measurements employing similar matrix-type sentence material. The perceived listening effort ratings further indicated that HI listeners had to put more effort into understanding the target speaker compared to NH listeners for both one and two interfering speakers, especially at lower SNRs.

An interactive graphic to compare the results of each measurement method between NH and HI listeners can be found at: `https://raginisinha.github.io/Interactive_comparisons/`.

Moreover, unlike NH listeners, HI listeners did not experience any reduction in effort at the lowest SNR (−10 dB) when the interfering speakers were of the opposite gender to the target speaker. Altogether, in line with previous studies [2], [225], [226], this study confirms that HI listeners struggle more in complex multi-talker scenarios compared to NH listeners and each considered evaluation method in this chapter is suitable to assess the performance of SC-TSE algorithms.

5.5.2 *Algorithm benefits for NH listeners vs. HI listeners*

For both NH and HI listeners, Algo-1 did not provide any improvement over unprocessed stimuli in terms of preference, speech intelligibility, or listening effort, while Algo-2 consistently demonstrated improvements over unprocessed stimuli for all considered evaluation methods. Notably, Algo-2 provided greater improvements for HI listeners compared to NH listeners for all evaluation methods. For NH listeners, Algo-2 improved mean SRTs (~ 3 dB) only for male interfering speakers compared to the unprocessed stimuli, while for HI listeners, Algo-2 showed improvements for both male (~ 3 dB) and female (~ 1 dB) interfering speakers. This could be because of the much higher SRTs for unprocessed stimuli with female interfering speakers for HI listeners (-5 to -6 dB) compared to NH listeners (-13 dB). When comparing benefits between NH and HI listeners, Algo-2 showed a similar benefit for both NH and HI listeners for male interfering speaker with no statistically significant difference according to pairwise contrast analysis using a linear mixed-effects model. In contrast, for the female interfering speakers, significant differences to NH listeners of 4.7 dB (aided) and 4.2 dB (unaided) were observed (aided: 95% CI $[2.75, 6.56]$, $p < .001$; unaided: 95% CI $[2.28, 6.09]$, $p < .001$). Algo-2 also showed a reduction in listening effort over a broad range of SNRs for both NH listeners and HI listeners for both one and two interfering speakers. However, the benefits were more pronounced for HI listeners with reductions of $7 - 8$ ESCU compared to $4 - 5$ ESCU for NH listeners at medium SNRs. Even at the lowest SNR (-10 dB) for one interfering speaker, the benefit for HI listeners ($4 - 5$ ESCU) was greater than for NH listeners ($1 - 2$ ESCU). Pairwise contrast analysis using a linear mixed-effects model between NH and HI listeners for the listening effort benefits also showed that HI listeners in both unaided and aided conditions showed significantly larger benefits compared to NH listeners at each SNR (all $p < 0.001$). The paired comparison results showed a similar trend for both NH listeners and HI listeners. However, HI listeners displayed a stronger preference for Algo-2, with the majority of their ratings falling into the "much easier" category of the rating scale. Across all evaluation methods, Algo-2 provided greater benefits for HI listeners compared to NH listeners.

Furthermore, we also conducted Pearson correlation analyses to assess whether hearing loss severity, as measured by PTA4 (averaged across both ears), is related to the SRTs and listening effort benefits of Algo-2, analyzed separately for male and female interfering speakers. No significant correlations were observed between hearing loss severity and algorithm benefit across any condition, unaided or aided, with one or two interfering speakers, or at any SNR.

5.5.3 *Impact of algorithmic artifacts for both NH and HI listeners*

In general, SC-TSE algorithms may introduce artifacts in the processed target speaker signals, mainly depending on the SNR of the mixture, the number of interfering speakers, and the used speaker separator network. Typical artifacts include distortions of the extracted target speaker and residual interference from the interfering speaker(s). As already mentioned, Algo-1 performs target speaker extraction in the time-frequency domain using a real-valued mask (hence using the mixture phase) with a relatively simple network architecture, whereas Algo-2 performs target speaker extraction in the time domain with a more complex network architecture. Several studies have shown that real-valued spectral masking-based approaches typically introduce more artifacts compared to time-domain approaches.

A significant impact of artifacts introduced by Algo-1 was observed for both NH and HI listeners. Algo-1 not only failed to reduce listening effort compared to unprocessed stimuli but, in some cases even increased the required effort. Specifically, for HI listeners, the listening effort was higher at 10 dB (female) and 15 dB (male) for one interfering speaker, and at 5 dB (female) and 15 dB (male) for two interfering speakers. These findings are inconsistent with objective assessments, where Algo-1 showed considerable improvement in target speaker quality and intelligibility. This difference arises because Algo-1 introduces artifacts that significantly impair speech perception for human listeners, but which may not be fully reflected by objective metrics. The artifacts introduced by Algo-1 had a more pronounced effect on HI listeners (without hearing loss compensation) than on NH listeners, as observed during SRT measurements. For male interfering speakers in the unaided condition, some HI listeners started to follow one of the interfering speakers instead of the target speaker, leading to invalid data. For HI listeners, it is particularly challenging to focus on the target speaker when the mixture includes interfering speakers of the same gender. Algorithm artifacts, such as residuals of the interfering speaker(s) in the processed signal, exacerbate this difficulty. Without hearing loss compensation, HI listeners are more likely to miss crucial features of the target speaker's voice that help distinguish it from interfering speakers.

5.5.4 *Effects of hearing loss compensation for HI listeners*

Apart from the appearance of such invalid data during the SRT measurements, this study found no significant differences between evaluations without (unaided) and with (aided) hearing loss compensation. Results from all three evaluation methods showed very similar patterns for both unaided and aided conditions. Even though SRTs for the aided condition (unprocessed, Algo-1 and Algo-2) were slightly lower than for the unaided condition, the difference was small (1.1 dB on average), and no statistically significant interactions between hearing loss compensation and other factors were found. The statistical analysis for listening effort benefits also confirmed this. Similar observations were reported by [227] for hearing loss compensation. Although linear amplification can make sounds audible for HI listeners in noisy environments, it does not significantly improve the ability to understand speech in

multi-talker scenarios. This study confirms that linear amplification alone is not sufficient to improve target speech perception (see results for unprocessed stimuli for all evaluation methods). Therefore, to tackle the challenges in cocktail party problem, target speaker extraction algorithms such as the ones investigated in this study are required, as they can provide substantial benefits with or without hearing loss compensation.

In other words, hearing loss compensation may not be needed for these target speaker extraction algorithms to be beneficial, especially when the participant has mild to moderate hearing loss. We found the algorithmic benefit to be independent of amplification, which applies only to the group tested in this study and may differ for individuals with more severe hearing loss. One possibly interesting implication of this observation is that SC-TSE algorithms could be a valuable future asset of hearables that target a broader user range than people with known hearing loss. This target group could benefit from assistive listening comprising speaker extraction even when increased hearing thresholds play a minor role in their daily life or when they are not aware of audibility impairments. We also observed that speech audibility was not a critical factor, as no substantial differences were found between the aided and unaided conditions.

5.5.5 *Limitations of this study*

One limitation of the study conducted in this chapter is that all participants had relatively symmetric hearing loss. It is possible that the role of hearing loss compensation could differ for participants with asymmetric hearing loss, where each ear requires a different level of amplification. Additionally, all participants had mild to moderate hearing loss, therefore, the effects of hearing loss compensation observed here may not generalize to individuals with more severe hearing loss, who may respond differently to SC-TSE algorithms.

The results also suggest that both NH and HI listeners could benefit significantly from SC-TSE algorithms. Especially for HI listeners, Algo-2 can bring considerable benefits if it is implemented in practical applications such as hearing aids. However, for an algorithm to be suitable for hearing aids, it needs to meet specific requirements. The algorithm needs to be capable of real-time processing, i.e., have low computational complexity and an algorithmic latency of 10 ms or less [228], [229]. The algorithm needs to meet the hardware constraints of the device, balancing performance and power consumption. Although Algo-2 is capable of real-time processing, it falls short in other critical aspects, such as memory size and computational complexity. Furthermore, this chapter focused only on the impact of one and two interfering speakers in the mixture signal. Other factors, such as background noise and reverberation, remain unexplored for both NH and HI listeners.

5.6 Summary

In this chapter, we conducted subjective evaluations of two different SC-TSE algorithms using both NH and HI listeners. For HI listeners, we considered both unaided

(without compensation) and aided (with linear hearing loss compensation) conditions. Three evaluation methods were used: paired comparison, speech recognition thresholds (SRTs), and categorically scaled perceived listening effort. Each method was applied to a group of 15 NH and 15 HI listeners to assess the effectiveness of both SC-TSE algorithms.

The evaluation results showed that Algo-2 consistently showed its benefits compared to the unprocessed mixture and Algo-1 across all evaluation methods for both NH and HI listeners. In contrast, Algo-1 did not provide any improvement compared to the unprocessed mixture. Furthermore, HI listeners experienced greater reductions in listening effort at lower SNRs and more improvements in SRTs than NH listeners. Algo-2 was consistently preferred by both NH and HI listeners over both the unprocessed mixture and Algo-1. Despite these advantages, processing artifacts introduced by the SC-TSE algorithms, particularly Algo-1, were perceived by both NH and HI listeners. The results also suggested that hearing loss compensation did not significantly affect the performance of the SC-TSE algorithms with the chosen participants in this chapter. Also, across all evaluation methods indicated no substantial difference between the unaided and aided conditions.

6

CONCLUSIONS AND FURTHER RESEARCH

In this chapter we summarize the main results of this thesis and indicate directions for further research.

6.1 Conclusions

Target speaker extraction aims at extracting a specific speaker from a complex acoustic scenario where multiple interfering speakers are present. This task becomes even more challenging in the presence of background noise and reverberation. Enhancing the intelligibility and quality of the target speaker in such scenarios is crucial for various applications. Traditionally, this problem has been approached through blind source separation, which aims to estimate all individual sound sources from the mixture. However, in many practical applications, extracting only the target speaker is sufficient, making full source separation unnecessary. Moreover, blind source separation algorithms require prior knowledge or estimation of the number of sources present in the mixture, which is often challenging. An alternative approach is speaker-conditioned target speaker extraction (SC-TSE), which directly estimates the target speaker by utilizing the auxiliary information about the target speaker (e.g., reference speech), mitigating the need to estimate the number of sources in the mixture.

The main objective of this thesis was to develop and evaluate novel DNN-based architectures leveraging reference speech of the target speaker to enhance the reliability, efficiency and robustness of single-channel SC-TSE algorithms. The first focus was to improve the speaker extraction performance and the real-time capability of SC-TSE algorithms using LSTM and conformer-based architectures. The second focus was to subjectively evaluate SC-TSE algorithms through listening tests with normal-hearing (NH) and hearing-impaired (HI) listeners to assess their real-world applicability.

As a first contribution, in Chapter 3 we proposed three novel architectures for the speaker separator network to perform target speaker extraction in the time-frequency domain. The customizations were introduced to enhance target speaker extraction by refining how information about the target speaker is retained and updated within the LSTM cells.

- Customized LSTM/BLSTM (F): this variant focuses on customizing only the forget gate (F) of the LSTM cells, aiming at selectively retaining target speaker features while disregarding interfering speakers and background noise present in the mixture.
- Customized LSTM/BLSTM (F+I): this variant focuses on customizing both the forget and input gates (F+I), aiming at refining speaker extraction performance further by improving the update mechanism of the cell state in LSTM, reinforcing the target speaker-specific characteristics retention.
- Customized auxiliary-gated LSTM/BLSTM: this variant introduces an additional modulation gate within the LSTM cell that dynamically enhances both long-term and short-term speaker feature discrimination.

Experimental results across various mixture types showed that each proposed variant improved target speaker extraction performance compared to standard LSTM cells. Among the proposed variants, the auxiliary-gated LSTM cells achieved the best performance, indicating that leveraging both long-term and short-term speaker feature discrimination dynamically through speaker embedding is highly effective in discriminating the target speaker from interfering sources. Results also showed that training on multi-condition mixtures enabled each system to generalize well to unseen test conditions, indicating that utilizing diverse mixture scenarios during training yields more robust SC-TSE systems. The generalization was particularly more evident in the case of noisy mixtures of three speakers (see Table 3.5). Furthermore, the customized LSTM (F+I) cells achieved significant performance improvement compared to standard LSTM cells with a substantial reduction (approximately 1.4 times) in the overall number of parameters (see Table 3.2). However, despite these advantages, some notable limitations were also observed. Specifically, for each proposed variant, the bidirectional mode did not achieve substantial performance improvements compared to the corresponding unidirectional mode, which suggests that access to future temporal context information provides limited additional information about the target speaker. Moreover, the bidirectional mode added extra computational costs compared to the unidirectional mode. For instance, the auxiliary-gated BLSTM increased the total number of parameters by approximately 1.8 times compared to its unidirectional mode (see Table 3.2) but showed no substantial performance improvement, which suggests a trade-off between the system complexity and performance improvement.

As a second contribution, in Chapter 4 we performed target speaker extraction in the time domain, as an alternative to the time-frequency domain-based target speaker extraction in Chapter 3. We proposed two different conformer-based architectures for the speaker separator network. Both architectures were designed to efficiently capture both local and global-context features by utilizing convolutional layers and multi-head self-attention (MHSA) in the conformer. The first proposed architecture (Conformer-FFN) consists of stacked conformer blocks and external feed-forward blocks, aiming at providing effective feature representation while reducing the overall number of parameters. The second proposed architecture (TCN-Conformer) consists of stacked TCN blocks and conformer blocks, aiming at exploiting the local-context features using TCN blocks and then exploiting both local and

global-context features using conformer blocks. Experimental results across various mixture types for varying the number of stacked blocks in both proposed architectures showed that the proposed TCN-Conformer system consistently outperformed both the TCN-based baseline system and the proposed Conformer-FFN systems. The best performance was achieved with TCN-Conformer, utilizing four stacks of TCN and conformer blocks. Furthermore, both proposed architectures showed systematic performance improvements with increasing number of stacked blocks. However, despite showing a benefit from an increased number of stacked blocks between conformer and FFN, the proposed Conformer-FFN system failed to outperform the TCN-based baseline system.

While the TCN-Conformer system shows strong speaker extraction performance, its reliance on traditional MHSA results in high computational and memory costs. We further extended our proposed TCN-Conformer system to make it more suitable for real-time target speaker extraction by replacing the traditional MHSA in each conformer block with linear MHSA. By introducing linear MHSA, the quadratic complexity of traditional MHSA was reduced to linear scaling with input signal length, significantly reducing both computational and memory costs. Additionally, a systematic reduction of the overall number of parameters of the system (Large, Medium, Small, XSmall) was investigated to further optimize real-time performance. Experimental results showed that the TCN-Conformer system using linear MHSA consistently outperformed TCN-Conformer systems using traditional MHSA while achieving a substantial reduction in computational cost and RTF. However, despite these advantages, some notable limitations were also observed. Although the Large variant of the proposed TCN-Conformer system with linear MHSA achieved the highest performance among all variants, it failed to meet the requirement for real-time processing (see Table 4.4).

Whereas in all previous experiments the reference speech was clean and matched during training and testing, in Appendix A we investigated the impact of mismatched reference speech on the performance of the TCN-Conformer system, proposed in Chapter 4. Our experiments revealed that the performance of the SC-TSE system degrades significantly when there is a mismatch between the training and testing conditions with regard to the reference speech of the target speaker, both due to external factors (noise and reverberation) as well as intrinsic speaker variability (emotions). This arises because mismatched reference speech leads to computations of ambiguous target speaker embedding, which may misguide the speaker separator network. To improve the robustness against these mismatches, we explored different strategies to use multi-condition training (MCT): either training the entire SC-TSE system from scratch with clean and augmented reference speech, or fine-tuning only the last layer of the speaker embedder network, or the speaker separator, or both networks. Experimental results showed that all utilized MCT strategies improved robustness against mismatched reference speech, with MCT from scratch showing the best performance across all test conditions. This was expected, as MCT from scratch allowed all layers in the SC-TSE system to learn from the entire data distribution, leading to more effective tuning of the parameters compared to the fine-tuning strategies.

While in the previous chapters 3 and 4, the performance of SC-TSE algorithms was only evaluated using objective measures, as a third contribution, in Chapter 5 we subjectively evaluated the performance of two SC-TSE algorithms by performing listening tests with NH and HI listeners (with and without hearing loss compensation). We considered two different SC-TSE algorithms: an algorithm performing target speaker extraction using a real-valued mask in the time-frequency domain (Algo-1) and an algorithm performing target speaker extraction in the time domain using the proposed TCN-Conformer architecture (Algo-2). Both algorithms were evaluated in challenging acoustic scenarios with up to two interfering speakers using three methods: paired comparison, speech recognition thresholds, and categorically scaled perceived listening effort. Both algorithms were evaluated using German matrix sentences from the Oldenburg Sentence Test (OLSA) for challenging acoustic scenarios with up to two interfering speakers. The interfering speakers, either male or female, were selected from a different matrix sentence dataset, while the target speaker was always a male speaker from OLSA.

In this study, we conducted each evaluation with fifteen NH and fifteen HI listeners. First, the evaluation results showed that all considered evaluation methods are suitable for evaluating SC-TSE algorithms. The evaluation results also showed a clear difference in how NH and HI listeners perceived the unprocessed mixture. HI listeners showed more difficulty in understanding the target speaker in a challenging acoustic scenario compared to NH listeners, especially at low SNRs. When comparing the effectiveness of Algo-1 and Algo-2, it was observed that Algo-2 consistently outperformed both the unprocessed mixtures and Algo-1 across all evaluation methods for both NH and HI listeners, where HI listeners experienced more benefits compared to NH listeners. For NH listeners, Algo-2 showed an improvement in SRT only for male interfering speakers compared to the unprocessed stimuli, while for HI listeners, it showed improvements for both male and female interfering speakers. When comparing SRT benefits between NH and HI listeners, Algo-2 showed a similar benefit for both NH and HI listeners for male interfering speakers, while for the female interfering speakers it showed a greater benefit for HI listeners compared to NH listeners. Additionally, in terms of listening effort benefits, Algo-2 showed that HI listeners experienced greater reduction in the perceived listening efforts at each SNR for both male and female interfering speakers compared to NH listeners. A similar benefit was also observed for paired comparisons, where HI listeners showed a stronger preference for Algo-2, with the majority of their ratings falling into the "much easier" category of the rating scale. In contrast, Algo-1 not only failed to provide any benefit over the unprocessed mixtures but also introduced artifacts that degraded the listening experience, especially for HI listeners. Furthermore, the results showed that the benefits obtained using Algo-2 were independent of hearing loss compensation, as no statistically significant differences were found between the evaluations performed without hearing loss compensation and with linear hearing loss compensation. However, it should be noted that all HI listeners who participated in this study had symmetric mild-to-moderate hearing loss. Therefore, the impact and requirement of hearing loss compensation may differ for individuals with more severe or asymmetric hearing loss.

6.2 Further research directions

In Chapter 3 we proposed three variants of LSTM cells to replace the standard LSTM cells in the speaker separator network of the time-frequency-domain SC-TSE algorithm. While each proposed variant outperformed standard LSTM cells, several aspects require further investigation.

First, a systematic investigation is required to gain a deeper understanding of how changes in the information flow through each gate influence the memory retention and speaker selectivity of the cell state. Specifically, how long the cell state retains the previous information and how quickly the hidden state responds to speaker changes. These investigations could not only help the redesigning of the information processing through each gate (as explored in Chapter 3) but also open possibilities for replacing the traditional gates with other DNN-based modules, similarly to [230].

Second, it remains unclear why the access to future temporal context information in bidirectional mode provided no substantial performance improvement compared to unidirectional mode for each proposed variant. It is unclear whether this is due to redundancy introduced by speaker embedding, saturation of global context within the chosen frame length, or inefficient use of future information. A systematic investigation into the contribution of each factor would be a valuable direction for future work. In this Chapter, we only considered real-valued mask estimation; future work could also explore the advantages of estimating the complex-valued masks using the proposed variants similarly to [82], [231].

In Chapter 4, we explored the benefits of exploiting both local and global context features utilizing conformer-based architectures to perform target speaker extraction in the time domain. We proposed two architectures, Conformer-FFN and TCN-Conformer. We further extended our best-performing TCN-Conformer to make it more suitable for real-time processing by replacing traditional MHSA with linear MHSA and systematically reducing the overall number of parameters.

Although linear MHSA was designed as a close approximation of traditional MHSA, it improved the speaker extraction performance compared to traditional MHSA. This improvement may be due to the ability of linear MHSA to generate smoother and more globally coherent temporal attention weights, which may help suppress artifacts more efficiently compared to traditional MHSA. A systematic investigation of gradient flow and attention weight distributions across layers for each SC-TSE system could be a valuable direction for future work. Such an investigation could provide insight into how attention weights are allocated and optimized across different heads during the training of the system, similar to [232]–[234], and could guide for further optimization for real-time applicability.

Furthermore, instead of using linear MHSA (as proposed in Chapter 4), other recent alternatives of MHSA [235], [236] could also be explored in future work to further enhance speaker extraction performance and real-time applicability. For instance, similarly to [23], the best-performing state-space-based Mamba architecture [236] could be combined with TCN or could be used to replace specific layers in each conformer block to enhance the target speaker extraction performance.

While Chapters 3 and 4 focused on utilizing only reference speech as auxiliary information to guide the speaker separator network, future research could focus on utilizing alternative or multi-modal auxiliary information to enhance target speaker extraction performance. Several studies have investigated alternative modalities [45], [48], [50], [51], [237], [238]. For instance, in [237] both reference speech and visual information were utilized, while in [45] reference speech, visual information, and spatial information were utilized, whereas in [50], [51] brain signals were utilized as auxiliary information to perform the target speaker extraction.

Since combining multiple auxiliary modalities has shown promising performance, especially when one modality is corrupted. One potential future direction could be combining brain signals of the listener with other modalities (e.g, reference speech, visual information, or spatial information), similar to [238]. Additionally, inspired by [172], future research could investigate the impact of combining different conditioning methods at different layers of the separator network. Another potential future research could be developing SC-TSE algorithms, which maintain the dynamic memory of the target speaker, enabling extraction even when the auxiliary information is corrupted, similarly to [239]. Throughout this thesis, we assumed that the target speaker was always active in the mixture. However, in a real-world environment, the target speaker may be temporarily inactive. Addressing this by incorporating techniques to detect and handle target speaker inactivity, as proposed in [145], could also be an important future direction. Furthermore, inspired by recent insights into system scalability and efficiency [240], [241], future research could focus on adapting large and complex speech separation and speech enhancement systems into lightweight, real-time SC-TSE systems utilizing knowledge distillation techniques.

In Chapters 3 and 4, we proposed different architectures for performing target speaker extraction either in the time-frequency domain or in the time domain, and evaluated their performance using objective measures. In Chapter 5, we conducted subjective evaluations through listening tests with both NH and HI listeners utilizing a time-frequency-based SC-TSE algorithm (Algo-1) and a time-domain-based SC-TSE algorithm (Algo-2). Across three different evaluation methods, Algo-2 consistently provided significant benefits for both NH and HI listeners compared to the unprocessed mixture, while Algo-1 showed no benefit. Interestingly, objective evaluations had previously shown that both Algo-1 and Algo-2 significantly improved target speaker extraction performance compared to the unprocessed mixture. This highlights a clear difference between objective and subjective evaluation results. To gain a deeper insight into it, future work could investigate the correlation between objective and subjective measures, similar to [242], [243].

A

STRATEGIES FOR ADDRESSING MISMATCHED REFERENCE SPEECH

While in all Chapters 3, 4, and 5, we considered clean reference speech for the training and testing of SC-TSE algorithms, in this Appendix, we investigate how mismatches in reference speech affect the performance of the SC-TSE algorithm. As discussed in Section 1.3.4, the performance of SC-TSE algorithms degrades in more realistic conditions where mismatches occur between training and testing conditions, for example, when the target and interfering speakers have similar voice characteristics [18] or when external factors like background noise and reverberation differ. Additionally, intrinsic speaker variability, such as differences in emotional states [80] can also impact the performance significantly.

We address the mismatch between the training and testing conditions in the previous Chapters 3 and 4 but only with regard to the mixture signal, i.e., how an unseen interfering speaker or background noise affected the target speaker extraction performance (see Section 3.4.2 and Section 4.4.1.1). In this Appendix, we focus on investigating the impact of the mismatches between the training and testing conditions with regard to the reference speech of the target speaker. A mismatch in reference speech during training and testing can lead the speaker embedder network to generate ambiguous target speaker embeddings [81], [82], which may confuse the speaker separator network in distinguishing the target speaker from the interfering speakers. To address this, we explore four different strategies based on multi-condition training (MCT) [244], [245] to improve the robustness and generalizability of the SC-TSE system proposed in Section 4.2.2.3. We mainly investigate the impact of mismatches due to external factors (noise and reverberation) and emotional states. Specifically, we explore four different MCT-based strategies: (1) training both the speaker embedder and separator networks from scratch, (2) fine-tuning the last layer of the speaker embedder network, (3) fine-tuning the last layer of the speaker separator network, and (4) fine-tuning the last layers of both speaker embedder and separator networks of the pre-trained SC-TSE system.

The effectiveness of the proposed MCT-based strategies is compared against the TCN-Conformer baseline system proposed in Section 4.2.2.3 (with 4 stacks of TCN and conformer blocks), which was trained with only clean reference speech.

The remainder of this Appendix is organized as follows: Section A.1 discusses each proposed MCT-based strategy to enhance the robustness of the SC-TSE system

against the mismatched reference speech. Section A.2 discusses the experimental setup, including the datasets, network architectures, and training and testing hyperparameters. Section A.3 discusses the experimental results and performance analysis of each considered system under different mismatches. Finally, Section A.4 provides a summary of this Appendix.

A.1 Target speaker extraction with mismatched reference speech

In a real-world scenario, it is highly likely that the reference speech of the target speaker $a_j(n)$ may have been recorded in a different acoustic environment or with a different emotional state of the target speaker as $x_j(n)$. As discussed in Section 4.1, the speaker embedder network generates the speaker embedding $\mathbf{e}_j = \boldsymbol{\phi}^{emb}(a_j(n))$ of the target speaker, which captures the target speaker-specific characteristics. The speaker separator network aims at extracting the target speaker signal $x_j(n)$ by jointly optimizing the speaker embedder network and speaker separator network, where the embedding guides the speaker separator network to distinguish the target speaker from the interfering speakers and background noise. However, if a mismatch occurs between the training and testing conditions in the reference speech, the speaker embedder network may generate an ambiguous speaker embedding, which confuses the speaker separator network [246] and may lead to the extraction of the wrong speaker. To address this, we explore MCT in four different ways (see Section A.1.1).

A.1.1 *Proposed MCT strategies for robust target speaker extraction*

We analyze the impact of noisy, reverberant, and emotional reference speech on target speaker extraction performance across different types of mixture, i.e., 2-speaker mixtures (2-mix), 3-speaker mixtures (3-mix), and noisy 2-speaker mixtures (noisy-mix) using TCN-Conformer system (see Section 4.2.2.3) as our baseline. Aiming to improve the performance of the baseline system for reference speech mismatches due to both external disturbances (background noise and reverberation) as well as intrinsic variability (emotions of the target speaker), we explore the benefits of utilizing MCT in four different ways.

- **MCT (scratch):** we utilize MCT to train the entire SC-TSE system from scratch, i.e., the weights and biases of both the speaker embedder and the speaker separator networks are randomly initialized. The system is trained using both clean and augmented (noisy, reverberant noisy, and emotional) reference speech for all types of mixtures (2-mix, 3-mix, and noisy-mix) (see Fig. A.1).

In general, learning from scratch can be costly in terms of time and resource consumption. Therefore, we also investigate the benefits of fine-tuning using MCT in the following ways:

- **Only embedder:** we fine-tune the last layer of the speaker embedder network (see Fig. A.2) while retraining the speaker separator network from scratch.

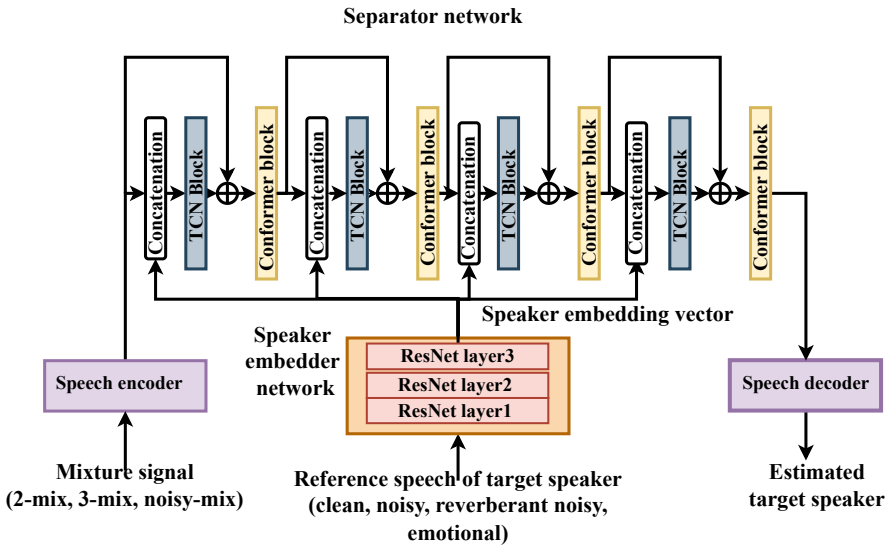

Fig. A.1: MCT strategy to train the SC-TSE system from scratch.

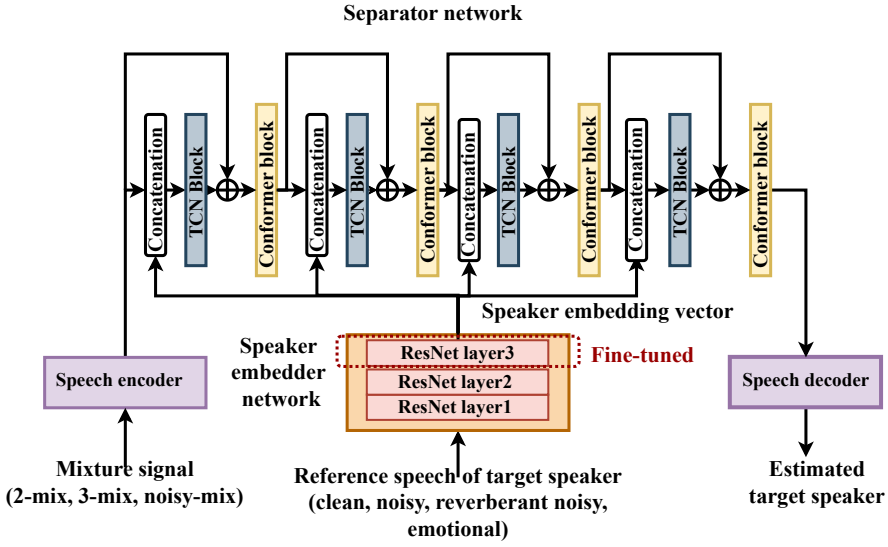

Fig. A.2: MCT strategy to fine-tune the last layer of the speaker embedder network of the pre-trained SC-TSE system.

- **Only separator:** we fine-tune the last layer of the speaker separator network (see Fig. A.3), while retraining the speaker embedder network from scratch.

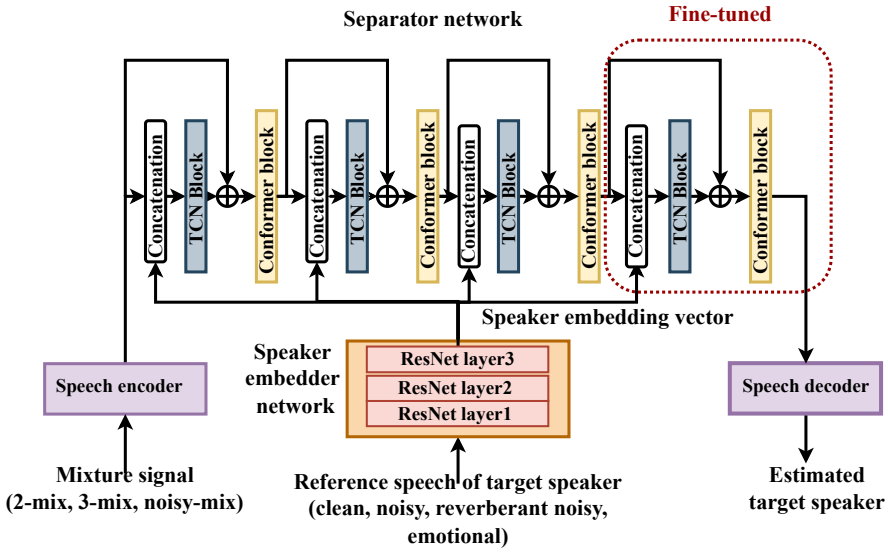

Fig. A.3: MCT strategy to fine-tune the last layer of the speaker separator network of the pre-trained SC-TSE system.

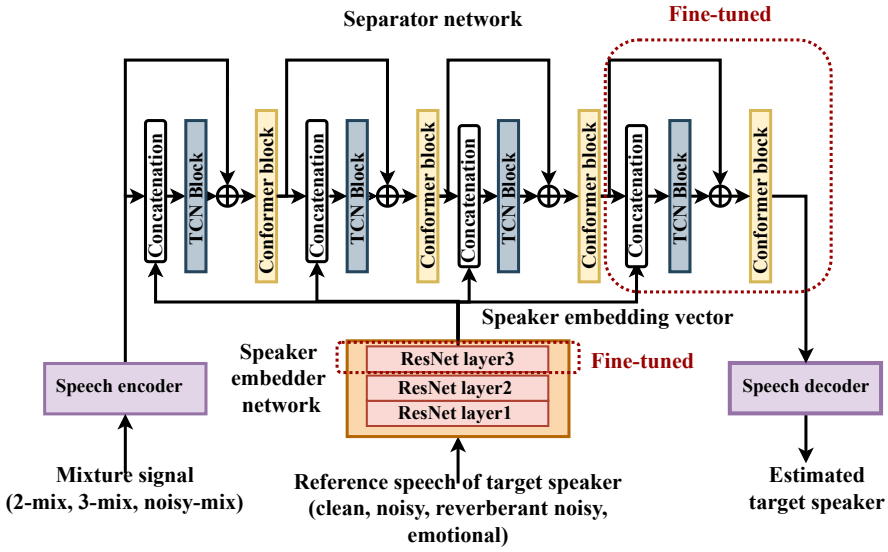

Fig. A.4: MCT strategy to fine-tune the last layer of both speaker embedder and separator networks of the pre-trained SC-TSE system.

- **Both:** we fine-tune the last layers of both speaker embedder and separator networks (see Fig. A.4).

For all proposed fine-tuning strategies, we use only augmented reference speech for all mixture types (2-mix, 3-mix, and noisy-mix), while the pre-trained baseline system was trained using only clean reference speech.

A.2 Experimental setup

In this section, we present the experimental setup used for all four systems discussed in Section A.1.1. In Section A.2.1, we discuss the datasets used for training and testing. In Section A.2.2, we discuss the training setup and hyperparameters used for each system.

A.2.1 *Datasets*

To generate the training, validation, and testing data, we considered four datasets with a sampling rate of 16 kHz, namely the WSJ0 dataset [137], the WHAM dataset [205], the emotional speech database (ESD) dataset [247] and the emotional RAVDESS2mix dataset [80]. The ESD dataset contains speech of 10 English and 10 Mandarin speakers for five emotions (neutral, happy, angry, surprised, sad). The RAVDESS2mix dataset contains speech of 24 different English speakers for eight emotions (neutral, happy, angry, surprised, sad, calm, disgust, fearful). The WSJ0 and WHAM datasets were used both for training/validation as well as for testing. The ESD dataset was only used for training/validation, whereas the RAVDESS2mix dataset was only used for testing. It should be noted that the testing data contains completely different speakers than the training and validation data.

A.2.1.1 *Training and validation data*

For training and validation, we have simulated 3 different types of mixtures: 2-mix, 3-mix, and noisy-mix, and four different types of reference speech: clean, noisy, reverberant noisy, and emotional. To create the mixtures, we followed the same procedure as discussed in Chapter 4. The baseline system is trained using only clean reference speech (we used the same system as mentioned in the last row of Table 4.3).

When performing MCT (both from scratch and fine-tuning), the same mixtures as for the baseline system were used, but besides clean reference speech, noisy reference speech and reverberant noisy reference speech were used. To create the noisy and reverberant noisy reference speech, noise samples were randomly chosen from the *tr* and *cv* subsets of the WHAM dataset and mixed with the clean reference speech at an SNR randomly chosen between -5 and 10 dB, while the RIRs were simulated using the image method [248], [249] with the same parameter settings as in [143]. In addition, training and validation data with emotional reference speech were simulated based on the official split of the ESD dataset. It should be noted that we only considered neutral English speech from the ESD dataset to simulate

2-mix and 3-mix, while the reference speech is a randomly chosen utterance of the target speaker with any of the four other emotions. To simulate 2-mix data for emotional reference speech, two different speakers are randomly chosen and mixed at an SNR between 0 and 5 dB. Similarly, for 3-mix data, both interfering speakers have the same power, and the mixture with the target speaker is simulated at an SNR between 0 and 5 dB. It should be noted that the simulated mixtures were anechoic for all types of reference speech.

A.2.1.2 *Testing data*

To evaluate target speaker extraction performance, we have performed two separate experiments: one addressing external acoustic conditions (noise and reverberation) and one addressing intrinsic speaker variability (emotions).

In the first experiment, we have evaluated the performance on 2-mix, 3-mix, and noisy-mix for three different types of reference speech (clean, noisy, and reverberant noisy). The mixtures were simulated as in Chapter 4 using the *si_dt_05* and *si_et_05* subsets of the WSJ0 dataset and the *tt* subset of the WHAM dataset. To create the noisy and reverberant noisy reference speech, noise samples were randomly chosen from the *tt* subset of the WHAM dataset and mixed with the clean reference speech at -5, 0, and 5 dB SNR, while the RIRs were simulated as in [143].

In the second experiment, we have evaluated the performance for emotional reference speech only on 2-mix test data. To ensure completely different test conditions in terms of both speakers and emotional states, we have used a completely different emotional (RAVDESS2mix) dataset than the dataset used for training and validation. It should be noted that the mixtures only contain the neutral speech of the target speaker, while the reference speech is a randomly chosen utterance of the target speaker with another emotion (happy, angry, surprised, sad, calm, disgust, fearful).

A.2.2 *Network architectures and training settings*

We have used the same TCN-Conformer SC-TSE system as discussed in Chapter 4 with the same hyperparameters as the baseline system. To incorporate MCT in the system, we have changed the learning rate as follows: during MCT from scratch, the system is trained with a learning rate of 0.001 using the ADAM optimizer [198], while during each proposed fine-tuning, the learning rate was reduced to 0.0001. All systems are trained using the same loss function with the same hyperparameters, a weighted combination of multi-scale SI-SDR loss for the separator network and cross-entropy loss for the embedder network (see Section 4.3.2).

A.3 Results and discussion

The performance of all considered SC-TSE systems was evaluated on test data described in Section A.2.1.2. As an evaluation measure, we have used the SI-SDR measure (dB) as discussed in Section 2.4.1. The clean target speaker signal was used

Test-set	**Baseline**	**MCT (scratch)**	**Only embedder**	**Only separator**	**Both**
	Clean reference speech				
2-mix	17.51	18.50	14.84	11.65	10.94
3-mix	10.70	12.24	7.34	6.04	5.15
Noisy-mix	9.32	10.51	2.83	1.88	-0.93
	Noisy reference speech				
2-mix	-18.92	**18.02**	14.72	12.48	11.25
3-mix	-19.03	**11.42**	7.16	7.14	5.11
Noisy-mix	-19.38	**10.44**	2.77	3.09	-1.14
	Reverberant noisy reference speech				
2-mix	-24.96	**17.10**	14.09	11.87	10.92
3-mix	-24.82	**11.21**	6.12	6.51	4.58
Noisy-mix	-24.58	**9.76**	0.76	2.62	-2.88

Table A.1: Mean SI-SDR (dB) for 2-mix, 3-mix, and noisy-mix for the baseline TCN-Conformer system and the proposed systems evaluated with clean reference speech, noisy reference speech, and reverberant noisy reference speech.

Test-set	Baseline	MCT (scratch)	Only embedder	Only separator	Both
Happy	-3.86	**10.72**	9.96	7.51	5.46
Angry	-3.31	**12.06**	10.96	7.34	6.34
Surprised	-7.40	**11.10**	10.14	7.17	5.67
Sad	-5.43	7.64	**9.51**	5.78	5.11
Calm	-4.43	8.44	**9.42**	5.88	5.02
Disgust	-3.91	**10.02**	9.75	6.17	6.01
Fearful	-4.49	**10.87**	10.21	7.28	6.18

Table A.2: Mean SI-SDR (dB) for 2-mix for the baseline system and the proposed systems evaluated with different types of emotional reference speech.

as the ground-truth signal. It should be noted that for all proposed fine-tunings, we use only augmented reference speech for all mixture types, while the pre-trained baseline system was trained using only clean reference speech.

Table A.1 shows the mean SI-SDR for different mixture types for the baseline system (trained using only clean reference speech) and the proposed systems, all evaluated with clean, noisy, and reverberant noisy reference speech. The results for the noisy and reverberant noisy reference speech have been averaged for SNRs -5, 0, and 5 dB in the reference speech signal. First, it can be observed that the performance of the baseline system degrades significantly for all mixture types when evaluated with noisy and reverberant noisy reference speech. Second, it can be observed that compared to the baseline system, each proposed MCT-based strategy significantly improves the performance for noisy and reverberant noisy reference speech. For all mixture types, the performance of the proposed MCT (scratch) evaluated with noisy and reverberant noisy reference speech is similar to the performance of the MCT (scratch) evaluated with clean reference speech and even similar to the performance of the baseline system evaluated with clean reference speech. Third, it can be observed that although each fine-tuning strategy yields better performance than the baseline system for all mixture types when evaluated with noisy and reverberant noisy reference speech, they yield worse performance than the baseline system when evaluated with clean reference speech and in general, MCT (scratch) outperforms each proposed fine-tuning.

Table A.2 shows mean SI-SDR for 2-mix for the baseline system (trained using only neutral speech) and the proposed systems, all evaluated with different types of emotional reference speech. First, it can be observed that the baseline system is highly sensitive to all types of emotions present in the reference speech. Second, it can be observed that compared to the baseline system, each proposed strategy significantly improves the performance for all types of emotions, including emotions not seen during training (calm, disgust, fearful). Except for sad and calm emotions, MCT (scratch) outperforms each fine-tuning.

A.4 Summary

In this Appendix, we investigated the impact of mismatched reference speech on the performance of a TCN-Conformer system and explored different strategies to enhance its robustness. We observed that mismatches in acoustic conditions (noise and reverberation) and intrinsic speaker variability (emotions) can significantly degrade the target speaker extraction performance of the system, especially when the considered system is trained using only clean reference speech. To address this, we explore the benefit of multi-condition training (MCT) in four different ways: training the SC-TSE system from scratch using clean and augmented reference speech, and fine-tuning (2) only the last layer of the speaker embedder, (3) only the last layer of the speaker separator, or (4) the last layers of both networks simultaneously of a pre-trained SC-TSE system.

Experimental results demonstrated that the baseline TCN-Conformer system, trained using only clean reference speech, struggles with mismatched reference conditions, leading to a significant drop in performance. In contrast, all utilized MCT strategies improved robustness against mismatched reference speech, with the MCT from scratch achieving the best results across all testing conditions for each type of mixture. While the utilized fine-tuning strategies also provided performance gains, they were less effective than MCT from scratch, particularly when evaluated with clean reference speech.

BIBLIOGRAPHY

[1] A. W. Bronkhorst, "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, 2015.

[2] B. G. Shinn-Cunningham and V. Best, "Selective Attention in Normal and Impaired Hearing," *Trends in Amplification*, vol. 12, no. 4, pp. 283–299, 2008.

[3] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. New York, USA: Springer, 2007.

[4] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. NJ, USA: John Wiley & Sons, 2018.

[5] 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 Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

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

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

[8] S. R. Chetupalli and E. A. Habets, "Speaker counting and separation from single-channel noisy mixtures," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 1681–1692, 2023.

[9] N. L. Westhausen and B. T. Meyer, "Binaural multichannel blind speaker separation with a causal low-latency and low-complexity approach," *IEEE Open J. of Signal Processing*, vol. 5, pp. 238–247, 2023.

[10] M. Elminshawi, S. R. Chetupalli, and E. A. Habets, "Slim-TasNet: A slimmable neural network for speech separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, NY, USA, Oct. 2023, pp. 1–5.

[11] M. Elminshawi, S. R. Chetupalli, and E. A. Habets, "Dynamic Slimmable Network for Speech Separation," *IEEE Signal Processing Letters*, pp. 2205–2209, 2024.

[12] S. Araki, N. Ito, R. Haeb-Umbach, G. Wichern, Z.-Q. Wang, and Y. Mitsufuji, "30+ years of source separation research: Achievements and future challenges," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr. 2025, pp. 1–5.

[13] K. Žmolíková, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[14] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE J. of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[15] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2728–2732.

[16] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-scale time domain speaker extraction network," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 1370–1384, 2020.

[17] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A complete time domain speaker extraction network," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 1406–1410.

[18] M. Delcroix, T. Ochiai, K. Žmolíková, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May. 2020, pp. 691–695.

[19] X. Li, R. Liu, H. Huang, and Q. Wu, "Contrastive learning for target speaker extraction with attention-based fusion," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 178–188, 2023.

[20] N. Kamo, M. Delcroix, and T. Nakatani, "Target speech extraction with conditional diffusion model," in *Proc. Interspeech*, Dublin, Ireland, Aug. 2023, pp. 176–180.

[21] J. Wang, H. Liu, L. Xu, W. Yang, W. Yi, and F. Liu, "Lightweight target speaker separation network based on joint training," *EURASIP J. on Audio, Speech, and Music Processing*, vol. 2023, no. 1, p. 53, 2023.

[22] J.-H. Wang, Y.-T. Lai, T.-C. Tai, P. T. Le, T. Pham, Z.-Y. Wang, Y.-H. Li, J.-C. Wang, and P.-C. Chang, "Target speaker extraction using attention-enhanced temporal convolutional network," *Electronics*, vol. 13, no. 2, p. 307, 2024.

[23] H. Sato, T. Moriya, M. Mimura, S. Horiguchi, T. Ochiai, T. Ashihara, A. Ando, K. Shinayama, and M. Delcroix, "Speakerbeam-ss: Real-time target speaker extraction with lightweight conv-tasnet and state space modeling," in *Proc. Interspeech*, Kos, Greece, Sept. 2024, pp. 5033–5037.

[24] J. Peng, M. Delcroix, T. Ochiai, O. Plchot, S. Araki, and J. Černockỳ, "Target speech extraction with pre-trained self-supervised learning models," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Apr. 2024, pp. 10 421–10 425.

[25] J. Peng, T. Ashihara, M. Delcroix, T. Ochiai, O. Plchot, S. Araki, and J. Černockỳ, "TS-SUPERB: A Target Speech Processing Benchmark for Speech Self-Supervised Learning Models," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr. 2025, pp. 1–5.

[26] J. Hershey and M. Casey, "Audio-visual sound separation via hidden markov models," *Advances in Neural Information Processing Systems*, vol. 14, 2001.

[27] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual Speech Source Separation: An overview of key methodologies," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 125–134, 2014.

[28] R. Lu, Z. Duan, and C. Zhang, "Audio–visual deep clustering for speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1697–1712, 2019.

[29] J. Lee, S.-W. Chung, S. Kim, H.-G. Kang, and K. Sohn, "Looking into your speech: Learning cross-modal affinity for audio-visual speech separation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, USA, June. 2021, pp. 1336–1345.

[30] T. Pan, J. Liu, B. Wang, J. Tang, and G. Wu, "RAVSS: Robust Audio-Visual Speech Separation in Multi-Speaker Scenarios with Missing Visual Cues," in *Proc. ACM International Conference on Multimedia*, Melbourne, Australia, Oct. 2024, pp. 4748–4756.

[31] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. on Graphics*, vol. 37, no. 4, pp. 1–11, 2018.

[32] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

[33] S. S. Shetu, S. Chakrabarty, and E. A. Habets, "An empirical study of visual features for DNN based audio-visual speech enhancement in multi-talker environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, June. 2021, pp. 8418–8422.

[34] Z. Pan, M. Ge, and H. Li, "USEV: Universal speaker extraction with visual cue," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 3032–3045, 2022.

[35] J. Li, M. Ge, Z. Pan, R. Cao, L. Wang, J. Dang, and S. Zhang, "Rethinking the visual cues in audio-visual speaker extraction," in *Proc. Interspeech*, Dublin, Ireland, Aug. 2023, pp. 3754–3758.

[36] M. Elminshawi, W. Mack, S. R. Chetupalli, S. Chakrabarty, and E. A. Habets, "New insights on the role of auxiliary information in target speaker extraction," *Frontiers in Signal Processing*, vol. 4, p. 1 440 401, 2024.

[37] R. Gu, L. Chen, S.-X. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 4290–4294.

[38] G. Li, S. Liang, S. Nie, W. Liu, M. Yu, L. Chen, S. Peng, and C. Li, "Direction-Aware Speaker Beam for Multi-Channel Speaker Extraction," in *Proc. Interspeech*, Graz, Austria, Sept. 2019, pp. 2713–2717.

[39] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "L-SpEx: Localized target speaker extraction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 2022, pp. 7287–7291.

[40] M. Elminshawi, S. R. Chetupalli, and E. A. Habets, "Beamformer-guided target speaker extraction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, June. 2023, pp. 1–5.

[41] S. He, J. Liu, H. Li, Y. Yang, F. Chen, and X. Zhang, "3S-TSE: Efficient three-stage target speaker extraction for real-time and low-resource applications," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Apr. 2024, pp. 421–425.

[42] M. Delcroix, K. Žmolíková, T. Ochiai, K. Kinoshita, and T. Nakatani, "Speaker activity driven neural speech extraction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, June 2021, pp. 6099–6103.

[43] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE J. of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 530–541, 2020.

[44] J. Han, X. Zhou, Y. Long, and Y. Li, "Multi-channel target speech extraction with channel decorrelation and target speaker adaptation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, June. 2021, pp. 6094–6098.

[45] J. Xu, J. Cui, Y. Hao, and B. Xu, "Multi-cue guided semi-supervised learning toward target speaker separation in real environments," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 151–163, 2023.

[46] M. Elminshawi, W. Mack, S. Chakrabarty, and E. A. Habets, *New insights on target speaker extraction*, Preprint at `https://arxiv.org/abs/2202.00733`, 2022.

[47] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, D. S. Williamson, and D. Yu, "Multi-channel multi-frame ADL-MVDR for target speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 3526–3540, 2021.

[48] B. Veluri, M. Itani, T. Chen, T. Yoshioka, and S. Gollakota, "Look once to hear: Target speech hearing with noisy examples," in *Proc. CHI Conference on Human Factors in Computing Systems*, HI, USA, May 2024, pp. 1–16.

[49] E. Ceolini, J. Hjortkjär, D. D. Wong, J. O'Sullivan, V. S. Raghavan, J. Herrero, A. D. Mehta, S.-C. Liu, and N. Mesgarani, "Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception," *NeuroImage*, vol. 223, p. 117 282, 2020.

[50] Z. Pan, M. Borsdorf, S. Cai, T. Schultz, and H. Li, "NeuroHeed: Neuro-Steered Speaker Extraction Using EEG Signals," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 4456–4470, 2024.

[51] D. De Silva, S. Cai, S. Pahuja, T. Schultz, and H. Li, "Neurospex: Neuro-Guided Speaker Extraction With Cross-Modal Fusion," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Macao, China, Dec. 2024, pp. 341–348.

[52] R. Togneri and D. Pullella, "An Overview of Speaker Identification: Accuracy and Robustness Issues," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 23–61, 2011.

[53] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 4879–4883.

[54] J. Malek, J. Jansky, Z. Koldovsky, T. Kounovsky, J. Cmejla, and J. Zdansky, "Target speech extraction: Independent vector extraction guided by supervised speaker identification," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 2295–2309, 2022.

[55] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC press, 2007.

[56] N. L. Westhausen and B. T. Meyer, "Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 2477–2481.

[57] N. L. Westhausen, H. Kayser, T. Jansen, and B. T. Meyer, "Real-time multi-channel deep speech enhancement in hearing aids: Comparing monaural and binaural processing in complex acoustic scenarios," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 4596–4606, 2024.

[58] D. O'Shaughnessy, "Speech enhancement - a review of modern methods," *IEEE Trans. on Human-Machine Systems*, vol. 54, no. 1, pp. 110–120, 2024.

[59] H. Wang, A. Pandey, and D. Wang, "A systematic study of DNN based speech enhancement in reverberant and reverberant-noisy environments," *Computer Speech & Language*, vol. 89, p. 101 677, 2025.

[60] Y. Xie and Z.-H. Tan, "A Survey of Deep Learning for Complex Speech Spectrograms," *arXiv preprint arXiv:2505.08694*, 2025.

[61] S. S. Shetu, E. A. Habets, and A. Brendel, "GAN-based Speech Enhancement for Low SNR Using Latent Feature Conditioning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr. 2025, pp. 1–5.

[62] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1670–1679, 2015.

[63] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Proc. International Conference on Signal Processing (ICSP)*, Hangzhou, China, Oct. 2014, pp. 473–477.

[64] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 967–977, 2016.

[65] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[66] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[67] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 274–278.

[68] C.-Y. Cheng, H.-S. Lee, Y. Tsao, and H.-M. Wang, "Multi-target extractor and detector for unknown-number speaker diarization," *IEEE Signal Processing Letters*, vol. 30, pp. 638–642, 2023.

[69] C. Boeddeker, A. S. Subramanian, G. Wichern, R. Haeb-Umbach, and J. Le Roux, "TS-SEP: Joint diarization and separation conditioned on estimated speaker embeddings," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 1185–1197, 2024.

[70] L. Serafini, S. Cornell, G. Morrone, E. Zovato, A. Brutti, and S. Squartini, "An experimental review of speaker diarization methods with application to two-speaker conversational telephone speech recordings," *Computer Speech & Language*, vol. 82, p. 101 534, 2023.

[71] M. Rybicka, J. Villalba, T. Thebaud, N. Dehak, and K. Kowalczyk, "End-to-end neural speaker diarization with non-autoregressive attractors," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2024.

[72] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Olomouc, Czech Republic, Dec. 2013, pp. 55–59.

[73] M. Karafiát, K. Veselỳ, J. Profant, J. Nytra, M. Hlaváček, and T. Pavlíček, "Analysis of X-Vectors for Low-Resource Speech Recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, June. 2021, pp. 6998–7002.

[74] H. Kuttruff, *Acoustics: an introduction*. CRC Press, 2007.

[75] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[76] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Springer Science & Business Media, 2010.

[77] G. Morrone, S. Cornell, E. Zovato, A. Brutti, and S. Squartini, "Conversational Speech Separation: an Evaluation Study for Streaming Applications," in *Proc. Audio Engineering Society Convention 152*, Hague, Netherlands, May 2022.

[78] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-Channel Multi-Speaker Separation Using Deep Clustering," in *Proc. Interspeech*, San Francisco, USA, Sept. 2016, pp. 545–549.

[79] D. Ditter and T. Gerkmann, "Influence of speaker-specific parameters on speech separation systems," in *Proc. Interspeech*, Graz, Austria, Sept. 2019, pp. 4584–4588.

[80] J. Švec, K. Žmolíková, M. Kocour, M. Delcroix, T. Ochiai, L. Mošner, and J. H. Černockỳ, "Analysis of impact of emotions on target speech extraction and speech separation," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sept. 2022, pp. 1–5.

[81] Z. Zhao, D. Yang, R. Gu, H. Zhang, and Y. Zou, "Target Confusion in End-to-end Speaker Extraction: Analysis and Approaches," in *Proc. Interspeech*, Incheon, Korea, Sept. 2022, pp. 5333–5337.

[82] Y. Liu, X. Liu, X. Miao, and J. Yamagishi, "Target speaker extraction with curriculum learning," in *Proc. Interspeech*, Kos, Greece, Sept. 2024, pp. 4348–4352.

[83] Y. Liu, X. Liu, X. Miao, and J. Yamagishi, "Libri2Vox Dataset: Target Speaker Extraction with Diverse Speaker Conditions and Synthetic Data," *arXiv preprint arXiv:2412.12512*, 2024.

[84] F.-L. Wang, H.-S. Lee, Y. Tsao, and H.-M. Wang, "Disentangling the Impacts of Language and Channel Variability on Speech Separation Networks," in *Proc. Interspeech*, Incheon, Korea, Sept. 2022, pp. 5343–5347.

[85] M. Borsdorf, C. Xu, H. Li, and T. Schultz, "Universal Speaker Extraction in the Presence and Absence of Target Speakers for Speech of One and Two Talkers," in *Proc. Interspeech*, Brno, Czechia, Aug. 2021, pp. 1469–1473.

[86] M. Delcroix, K. Kinoshita, T. Ochiai, K. Žmolíková, H. Sato, and T. Nakatani, "Listen only to me! How well can target speech extraction handle false alarms?" In *Proc. Interspeech*, Incheon, Korea, Sept. 2022, pp. 216–220.

[87] S. R. Chetupalli and E. A. Habets, "A Unified Approach to Speaker Separation and Target Speaker Extraction Using Encoder-Decoder Based Attractors," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aalborg, Denmark, Sept. 2024, pp. 190–194.

[88] E. Rykova and S. Werner, "Perceptual and acoustic analysis of voice similarities between parents and young children," in *Proc. Nordic Conference on Computational Linguistics*, Turku, Finland, Sept. 2019, pp. 262–271.

[89] S. Taylor, C. Dromey, S. L. Nissen, K. Tanner, D. Eggett, and K. Corbin-Lewis, "Age-related changes in speech and voice: Spectral and cepstral measures," *J. of Speech, Language, and Hearing Research*, vol. 63, no. 3, pp. 647–660, 2020.

[90] E. Z. Golumbic, G. B. Cogan, C. E. Schroeder, and D. Poeppel, "Visual input enhances selective speech envelope tracking in auditory cortex at a cocktail party," *J. of Neuroscience*, vol. 33, no. 4, pp. 1417–1426, 2013.

[91] H. Sato, T. Ochiai, K. Kinoshita, M. Delcroix, T. Nakatani, and S. Araki, "Multimodal attention fusion for target speaker extraction," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, Jan. 2021, pp. 778–784.

[92] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "Multimodal SpeakerBeam: Single Channel Target Speech Extraction with Audio-Visual Speaker Clues," in *Proc. Interspeech*, Graz, Austria, Sept. 2019, pp. 2718–2722.

[93] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "Multimodal speakerbeam: Single channel target speech extraction with audio-visual speaker clues," in *Proc. Interspeech*, Graz, Austria, 2019, pp. 2718–2722.

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

[95] Z. Pan, W. Wang, M. Borsdorf, and H. Li, "ImagineNet: Target speaker extraction with intermittent visual cue through embedding inpainting," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, June. 2023, pp. 1–5.

[96] S. Korse, M. Elminshawi, E. A. Habets, and S. R. Chetupalli, "Training strategies for modality dropout resilient multi-modal target speaker extraction," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, Seoul, Korea, Apr. 2024, pp. 595–599.

[97] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, "FaceFilter: Audio-Visual Speech Separation Using Still Images," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 3481–3485.

[98] J. A. O'sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial eeg," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[99] L. Straetmans, K. Adiloglu, and S. Debener, "Neural speech tracking and auditory attention decoding in everyday life," *Frontiers in Human Neuroscience*, vol. 18, p. 1 483 024, 2024.

[100] Y. Zhu, H. Li, Y. Liao, B. Wang, Z. Guan, H. Liu, and D. Cai, "What to do next: Modeling user behaviors by Time-LSTM.," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, VIC, vol. 17, Melbourne, Australia, Aug. 2017, pp. 3602–3608.

[101] R. Sinha, M. Tammen, C. Rollwage, and S. Doclo, "Speaker-conditioned target speaker extraction based on customized LSTM cells," in *Proc. ITG Conference on Speech Communication*, VDE, Kiel, Germany, Sept-Oct. 2021, pp. 1–5.

[102] R. Sinha, C. Rollwage, and S. Doclo, "Variants of LSTM cells for single-channel speaker-conditioned target speaker extraction," *EURASIP J. on Audio, Speech, and Music Processing*, vol. 2024, no. 1, pp. 1–13, 2024.

[103] R. Sinha, M. Tammen, C. Rollwage, and S. Doclo, "Speaker-conditioning single-channel target speaker extraction using conformer-based architectures," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sept. 2022, pp. 1–5.

[104] R. Sinha, C. Rollwage, and S. Doclo, "Real-time Single-channel Speaker-conditioned Target Speaker Extraction using TCN-Conformer with Efficient Self-attention Mechanisms," in *Proc. European Signal Processing Conference (EUSIPCO)*, Palermo, Italy, Sept. 2025, pp. 1–5.

[105] R. Sinha, A.-C. Scherer, S. Doclo, C. Rollwage, and J. Rennies, "Subjective performance evaluation of single-channel speaker-conditioned target speaker extraction algorithms for complex acoustic scenes," in *Proc. ITG Conference on Speech Communication*, Aachen, Germany, Sept. 2023, pp. 1–5.

[106] R. Sinha, A.-C. Scherer, S. Doclo, C. Rollwage, and J. Rennies, "Evaluation of speaker-conditioned target speaker extraction algorithms for hearing-impaired listeners," *Trends in Hearing*, vol. 29, p. 23 312 165 251 365 802, 2025.

[107] J. Flanagan, J. Johnston, R. Zahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78, no. 5, pp. 1508–1518, 1985.

[108] 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.

[109] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2002.

[110] L. Wang, H. Ding, and F. Yin, "Target speech extraction in cocktail party by combining beamforming and blind source separation.," *Acoustics Australia*, vol. 39, no. 2, 2011.

[111] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[112] Z. Chen, J. Li, X. Xiao, T. Yoshioka, H. Wang, Z. Wang, and Y. Gong, "Cracking the cocktail party problem by multi-beam deep attractor network," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, Dec. 2017, pp. 437–444.

[113] Z. Chen, T. Yoshioka, X. Xiao, L. Li, M. L. Seltzer, and Y. Gong, "Efficient integration of fixed beamformers and speech separation networks for multi-channel far-field speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 5384–5388.

[114] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, 2009.

[115] H. Chen, P. Zhang, and Y. Yan, "Multi-talker MVDR beamforming based on extended complex Gaussian mixture model," *arXiv preprint arXiv:1910.07753*, 2019.

[116] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 2655–2659.

[117] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning MVDR beamformer for target speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, June. 2021, pp. 6089–6093.

[118] A. Guo, J. Wu, P. Gao, W. Zhu, Q. Guo, D. Gao, and Y. Wang, "Enhanced Neural Beamformer with Spatial Information for Target Speech Extraction," in *Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Taipei, Taiwan, Nov. 2023, pp. 107–113.

[119] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF," *APSIPA Trans. on Signal and Information Processing*, vol. 8, e12, 2019.

[120] H. Sawada, R. Ikeshita, K. Kinoshita, and T. Nakatani, "Multi-frame full-rank spatial covariance analysis for underdetermined blind source separation and dereverberation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 3589–3602, 2023.

[121] C. A. Musluoglu and A. Bertrand, "Distributed Blind Source Separation based on FastICA," *IEEE Signal Processing Letters*, 2025.

[122] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[123] G. W. Taylor, M. L. Seltzer, and A. Acero, "Maximum a posteriori ICA: Applying prior knowledge to the separation of acoustic sources," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, USA, Apr. 2008, pp. 1821–1824.

[124] T.-W. Lee, "Independent component analysis," in *Independent component analysis: Theory and applications*, Springer, 1998, pp. 27–66.

[125] A. Hiroe, "Similarity-and-independence-aware beamformer with iterative casting and boost start for target source extraction using reference," *IEEE Open J. of Signal Processing*, vol. 3, pp. 1–20, 2021.

[126] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. International Conference on Independent Component Analysis and Signal Separation*, Springer, Charleston, USA, Mar. 2006, pp. 165–172.

[127] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Trans. on Signal Processing*, vol. 68, pp. 3545–3558, 2020.

[128] T. Nakashima and N. Ono, "Inverse-free online independent vector analysis with flexible iterative source steering," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Chiang Mai, Thailand, Nov. 2022, pp. 749–753.

[129] R. Scheibler, W. Zhang, X. Chang, S. Watanabe, and Y. Qian, "End-to-end multi-speaker ASR with independent vector analysis," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Doha, Qatar, Jan. 2023, pp. 496–501.

[130] J. Janský, J. Málek, J. Čmejla, T. Kounovský, Z. Koldovský, and J. Žd'ánský, "Adaptive blind audio source extraction supervised by dominant speaker identification using x-vectors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May. 2020, pp. 676–680.

[131] R. Scheibler and N. Ono, "Fast independent vector extraction by iterative SINR maximization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May. 2020, pp. 601–605.

[132] J. Janskỳ, Z. Koldovskỳ, J. Málek, T. Kounovskỳ, and J. Čmejla, "Auxiliary function-based algorithm for blind extraction of a moving speaker," *EURASIP J. on Audio, Speech, and Music Processing*, vol. 2022, no. 1, p. 1, 2022.

[133] T. Ueda, T. Nakatani, R. Ikeshita, S. Araki, and S. Makino, "DOA-informed switching independent vector extraction and beamforming for speech enhancement in underdetermined situations," *EURASIP J. on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 52, 2024.

[134] C. C. Aggarwal *et al.*, *Neural Networks and Deep Learning*. Springer, 2018, vol. 10.

[135] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[136] Y. Wang, J. Han, T. Zhang, and D. Qing, "Speech enhancement from fused features based on deep neural network and gated recurrent unit network," *EURASIP J. on Advances in Signal Processing*, vol. 2021, no. 1, p. 104, 2021.

[137] 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)*, Shanghai, China, Mar. 2016, pp. 31–35.

[138] J. Li, K. Zhang, S. Wang, H. Li, M.-W. Mak, and K. A. Lee, "On the effectiveness of enrollment speech augmentation for Target Speaker Extraction," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Macao, China, Dec. 2024, pp. 325–332.

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

[140] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 5329–5333.

[141] P. Matejka, O. Plchot, O. Glembek, L. Burget, J. Rohdin, H. Zeinali, L. Mošner, A. Silnova, O. Novotnỳ, and M. Diez, "13 years of speaker recognition research at BUT, with longitudinal analysis of NIST SRE," *Computer Speech & Language*, vol. 63, p. 101 035, 2020.

[142] C. Xu, W. Rao, E. S. Chng, and H. Li, "Time-domain speaker extraction network," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Sentosa, Singapore, Dec. 2019, pp. 327–334.

[143] M. Elminshawi, W. Mack, and E. A. Habets, "Informed source extraction with application to acoustic echo reduction," in *Proc. ITG Conference on Speech Communication*, VDE, Kiel, Germany, Sept-Oct. 2021, pp. 1–5.

[144] W. Liu and C. Xie, "Gated Convolutional Fusion for Time-Domain Target Speaker Extraction Network," in *Proc. Interspeech*, Incheon, Korea, Sept. 2022, pp. 5368–5372.

[145] S. Xu, Y. Yang, N. Trigoni, and A. Markham, "Target Speaker Extraction through Comparing Noisy Positive and Negative Audio Enrollments," *arXiv preprint arXiv:2502.16611*, 2025.

[146] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech*, Hyderabad, India, Sept. 2018, pp. 2252–2256.

[147] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*. MIT press Cambridge, 2016, vol. 1.

[148] Y.-Q. Yu, L. Fan, and W.-J. Li, "Ensemble additive margin softmax for speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May. 2019, pp. 6046–6050.

[149] D. Zhou, L. Wang, K. A. Lee, Y. Wu, M. Liu, J. Dang, and J. Wei, "Dynamic margin softmax loss for speaker verification," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 3800–3804.

[150] H. Bredin, "Tristounet: Triplet loss for speaker turn embedding," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, Mar. 2017, pp. 5430–5434.

[151] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 2655–2659.

[152] J. Heitkaemper, T. Fehér, M. Freitag, and R. Haeb-Umbach, "A study on online source extraction in the presence of changing speaker positions," in *Proc. International Conference on Statistical Language and Speech Processing*, Springer, Ljubljana, Slovenia, Oct. 2019, pp. 198–209.

[153] Y. Hao, J. Xu, J. Shi, P. Zhang, L. Qin, and B. Xu, "A Unified Framework for Low-Latency Speaker Extraction in Cocktail Party Environments.," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 1431–1435.

[154] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 32, New Orleans, USA, Feb. 2018.

[155] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan, C. Liu, D. Dimitriadis, J. Droppo, and Y. Gong, "Single-channel speech extraction using speaker inventory and attention network," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May. 2019, pp. 86–90.

[156] T. Li, Q. Lin, Y. Bao, and M. Li, "Atss-Net: Target speaker separation via attention-based neural network," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 1411–1415.

[157] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM trans. on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2015.

[158] F. Hao, A. Li, X. Li, and C. Zheng, "DSINet: Towards Real-Time Target Speaker Extraction with Dynamic Speaker Information Fusion," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr. 2025, pp. 1–5.

[159] K. Wilson, M. Chinen, J. Thorpe, B. Patton, J. Hershey, R. A. Saurous, J. Skoglund, and R. F. Lyon, "Exploring tradeoffs in models for low-latency speech enhancement," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sept. 2018, pp. 366–370.

[160] M. Tammen and S. Doclo, "Imposing Correlation Structures for Deep Binaural Spatio-Temporal Wiener Filtering," *IEEE Trans. on Audio, Speech and Language Processing*, 2025.

[161] 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)*, Brighton, UK, May 2019, pp. 626–630.

[162] Y. Koyama, T. Vuong, S. Uhlich, and B. Raj, "Exploring the best loss function for DNN-based low-latency speech enhancement with temporal convolutional networks," *arXiv:2005.11611*, 2020.

[163] 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.

[164] Z. Xu, M. Strake, and T. Fingscheidt, "Deep noise suppression maximizing non-differentiable PESQ mediated by a non-intrusive PESQNet," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 1572–1585, 2022.

[165] D. de Oliveira, S. Welker, J. Richter, and T. Gerkmann, "The PESQetarian: On the Relevance of Goodhart's Law for Speech Enhancement," in *Proc. Interspeech*, Kos, Greece, Sept. 2024, pp. 3854–3858.

[166] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs P.862," International Telecommunications Union (ITU-T) Recommendation, Tech. Rep., Feb. 2001.

[167] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[168] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *Proc. IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, Boston, USA, Aug. 2017, pp. 1597–1600.

[169] 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)*, Barcelona, Spain, May. 2020, pp. 46–50.

[170] S. Han, J. Byun, and J. W. Shin, "Time-domain speaker verification using temporal convolutional networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, June. 2021, pp. 6688–6692.

[171] N. N. An, N. Q. Thanh, and Y. Liu, "Deep CNNs with self-attention for speaker identification," *IEEE Access*, vol. 7, pp. 85 327–85 337, 2019.

[172] S. He, H. Zhang, W. Rao, K. Zhang, Y. Ju, Y. Yang, and X. Zhang, "Hierarchical speaker representation for target speaker extraction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Apr. 2024, pp. 10 361–10 365.

[173] K. Zhang, J. Li, S. Wang, Y. Wei, Y. Wang, Y. Wang, and H. Li, "Multi-level speaker representation for target speaker extraction," *arXiv:2410.16059*, 2024.

[174] W. Li, P. Zhang, and Y. Yan, "TEnet: target speaker extraction network with accumulated speaker embedding for automatic speech recognition," *Electronics Letters*, vol. 55, no. 14, pp. 816–819, 2019.

[175] M. Delcroix, K. Žmolíková, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 5554–5558.

[176] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Learning speaker representation for neural network based multichannel speaker extraction," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, Dec. 2017, pp. 8–15.

[177] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, T. Nakatani, and J. Černockỳ, "Optimization of speaker-aware multichannel speech extraction with ASR criterion," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 6702–6706.

[178] Q. Wang, I. L. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He, W. Li, J. Pelecanos, M. Nika, and A. Gruenstein, "VoiceFilter-Lite: Streaming Targeted Voice Separation for On-Device Speech Recognition," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 2677–2681.

[179] R. Rikhye, Q. Wang, Q. Liang, Y. He, D. Zhao, Y. Huang, A. Narayanan, and I. McGraw, "Personalized keyphrase detection using speaker and environment information," in *Proc. Interspeech*, Brno, Czechia, Aug-Sept. 2021, pp. 4204–4208.

[180] R. Rikhye, Q. Wang, Q. Liang, Y. He, and I. McGraw, "Multi-user VoiceFilter-Lite via attentive speaker embedding," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Cartagena, Colombia, Dec. 2021, pp. 275–282.

[181] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 6990–6994.

[182] J. Wang, J. Chen, D. Su, L. Chen, M. Yu, Y. Qian, and D. Yu, "Deep extractor network for target speaker recovery from single channel speech mixtures," in *Proc. Interspeech*, Hyderabad, India, Sept. 2018, pp. 307–311.

[183] 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)*, New Orleans, USA, Mar. 2017, pp. 246–250.

[184] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

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

[186] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2007.

[187] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, June. 2021, pp. 6493–6497.

[188] E. Parizet, "Paired comparison listening tests and circular error rates," *Acta acustica united with Acustica*, vol. 88, no. 4, pp. 594–598, 2002.

[189] K. Wagener, T. Brand, and B. Kollmeier, "Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil III : Evaluation des Oldenburger Satztests (Development and evaluation of a German sentence test Part III : Evaluation of the Oldenburg sentence test)," *Zeitschrift für Audiologie*, vol. 38, no. 3, pp. 86–95, 1999.

[190] J. Rennies, H. Schepker, I. Holube, and B. Kollmeier, "Listening effort and speech intelligibility in listening situations affected by noise and reverberation," *J. Acoust. Soc. Am.*, vol. 136, no. 5, pp. 2642–2653, 2014.

[191] W. Rao, C. Xu, E. S. Chng, and H. Li, "Target speaker extraction for multi-talker speaker verification," in *Proc. Interspeech*, Graz, Austria, 2019, pp. 1273–1277.

[192] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[193] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132 306, 2020.

[194] D. Neil, M. Pfeiffer, and S.-C. Liu, "Phased LSTM: Accelerating recurrent network training for long or event-based sequences," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[195] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 2616–2620.

[196] 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)*, Brisbane, Australia, Apr. 2015, pp. 5206–5210.

[197] Z. Zhang, B. He, and Z. Zhang, "X-TaSNet: Robust and Accurate Time-Domain Speaker Extraction Network," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 1421–1425.

[198] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference for Learning Representations*, San Diego, USA, Jul. 2015, pp. 1–15.

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

[200] D. Snyder, G. Chen, and D. Povey, *Musan: A music, speech, and noise corpus*, Preprint at `https://www.openslr.org/17/`, 2015.

[201] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 5036–5040.

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

[203] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T.-y. Liu, "Understanding and Improving Transformer From a Multi-Particle Dynamic System Point of View," in *Proc. ICLR Workshop on Integration of Deep Neural Models and Differential Equations*, Virtual, Apr. 2020, pp. 1–11.

[204] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv:1710.05941*, 2017.

[205] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "WHAM!: Extending Speech Separation to Noisy Environments," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 1368–1372.

[206] J. Lin, X. Cai, H. Dinkel, J. Chen, Z. Yan, Y. Wang, J. Zhang, Z. Wu, Y. Wang, and H. Meng, "Av-sepformer: Cross-attention sepformer for audio-visual target speaker extraction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, June 2023, pp. 1–5.

[207] Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li, "Efficient attention: Attention with linear complexities," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, HI, USA, Jan. 2021, pp. 3531–3539.

[208] M. N. Rabe and C. Staats, "Self-attention does not need o(n^2) memory," *arXiv:2112.05682*, 2021.

[209] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers," *J. Acoust. Soc. Am.*, vol. 125, no. 6, pp. 4006–4022, 2009.

[210] G. Kidd Jr, C. R. Mason, J. Swaminathan, E. Roverud, K. K. Clayton, and V. Best, "Determining the energetic and informational components of speech-on-speech masking," *J. Acoust. Soc. Am.*, vol. 140, no. 1, pp. 132–144, 2016.

[211] S. P. Bacon, J. M. Opie, and D. Y. Montoya, "The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds," *J. of Speech, Language, and Hearing Research*, vol. 41, no. 3, pp. 549–563, 1998.

[212] I. Reinten, I. De Ronde-Brons, R. Houben, and W. Dreschler, "Measuring the influence of noise reduction on listening effort in hearing-impaired listeners using response times to an arithmetic task in noise," *Trends in Hearing*, vol. 25, p. 23 312 165 211 014 437, 2021.

[213] I. Thoidis and T. Goehring, "Using deep learning to improve the intelligibility of a target speaker in noisy multi-talker environments for people with normal hearing and hearing loss," *J. Acoust. Soc. Am.*, vol. 156, no. 1, pp. 706–724, 2024.

[214] N. Bisgaard, M. S. Vlaming, and M. Dahlquist, "Standard audiograms for the IEC 60118-15 measurement procedure," *Trends in Amplification*, vol. 14, no. 2, pp. 113–120, 2010.

[215] B. Kollmeier and M. Wesselkamp, "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment," *J. Acoust. Soc. Am.*, vol. 102, no. 4, pp. 2412–2421, 1997.

[216] S. Hochmuth, B. Kollmeier, and B. Shinn-Cunningham, "The relation between acoustic-phonetic properties and speech intelligibility in noise across languages and talkers," in *Proc. Conf. Acoust. DAGA*, Munich, Germany, Mar. 2018, pp. 628–629.

[217] J. Rennies, V. Best, E. Roverud, and G. Kidd Jr., "Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort," *Trends in Hearing*, vol. 23, pp. 1–21, 2019.

[218] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Release from informational masking by time reversal of native and non-native interfering speech," *J. Acoust. Soc. Am.*, vol. 118, no. 3, pp. 1274–1277, 2005.

[219] K. Smeds, F. Wolters, and M. Rung, "Estimation of signal-to-noise ratios in realistic sound scenarios," *J. of the American Academy of Audiology*, vol. 26, no. 02, pp. 183–196, 2015.

[220] T. Brand and V. Hohmann, "An adaptive procedure for categorical loudness scaling," *J. Acoust. Soc. Am.*, vol. 112, no. 4, pp. 1597–1604, 2002.

[221] H. Dillon, *Hearing aids*. Thieme Medical Publishers, 2012.

[222] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. of Statistical Software*, vol. 67, no. i01, 2015.

[223] D. Lüdecke, M. S. Ben-Shachar, I. Patil, P. Waggoner, and D. Makowski, "Performance: An R package for assessment, comparison and testing of statistical models," *J. of Open Source Software*, vol. 6, no. 60, 2021.

[224] D. Makowski, M. S. Ben-Shachar, I. Patil, and D. Lüdecke, "Estimation of model-based predictions, contrasts and means," *CRAN*, 2020.

[225] G. Kidd, C. R. Mason, V. Best, E. Roverud, J. Swaminathan, T. Jennings, K. Clayton, and H. Steven Colburn, "Determining the energetic and informational components of speech-on-speech masking in listeners with sensorineural hearing loss," *J. Acoust. Soc. Am.*, vol. 145, no. 1, pp. 440–457, 2019.

[226] S. Hygge, J. Ronnberg, B. Larsby, and S. Arlinger, "Normal-hearing and hearing-impaired subjects' ability to just follow conversation in competing speech, reversed speech, and noise backgrounds," *J. of Speech, Language, and Hearing Research*, vol. 35, no. 1, pp. 208–215, 1992.

[227] B. Ohlenforst, A. A. Zekveld, E. P. Jansma, Y. Wang, G. Naylor, A. Lorens, T. Lunner, and S. E. Kramer, "Effects of hearing impairment and hearing aid amplification on listening effort: A systematic review," *Ear and hearing*, vol. 38, no. 3, pp. 267–281, 2017.

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

[229] L. Bramsløw, "Preferred signal path delay and high-pass cut-off in open fittings," *International J. of Audiology*, vol. 49, no. 9, pp. 634–644, 2010.

[230] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Fully convolutional recurrent networks for speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May. 2020, pp. 6674–6678.

[231] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, "DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shenzhen, China, May. 2022, pp. 7407–7411.

[232] L. Liu and X. Xu, "Self-attention mechanism at the token level: Gradient analysis and algorithm optimization," *Knowledge-Based Systems*, vol. 277, p. 110 784, 2023.

[233] Y. Liang, J. Long, Z. Shi, Z. Song, and Y. Zhou, "Beyond linear approximations: A novel pruning approach for attention matrix," *arXiv preprint arXiv:2410.11261*, 2024.

[234] C. Siyu, S. Heejune, W. Tianhao, and Y. Zhuoran, "Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality," in *Proc. Annual Conference on Learning Theory*, Edmonton, Canada, June. 2024, pp. 4573–4573.

[235] L. Liu, L. Cai, C. Zhang, X. Zhao, J. Gao, W. Wang, Y. Lv, W. Fan, Y. Wang, M. He, *et al.*, "Linrec: Linear attention mechanism for long-term sequential recommender systems," in *Proc. International ACM SIGIR Conference on Research and Development in Information Retrieval*, Taipei, Taiwan, July. 2023, pp. 289–299.

[236] R. Whetten, T. Parcollet, A. Moumen, M. Dinarelli, and Y. Estève, "An analysis of linear complexity attention substitutes with best-rq," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Macau, China, Dec. 2024, pp. 169–176.

[237] Z. Mu and X. Yang, "Separate in the speech chain: Cross-modal conditional audio-visual target speech extraction," *arXiv preprint arXiv:2404.12725*, 2024.

[238] C. Fan, Y. Chen, J. Zhou, Z. Pan, J. Zhang, Y. Gao, X. Yang, Z. Wen, and Z. Lv, "M3ANet: Multi-scale and Multi-Modal Alignment Network for Brain-Assisted Target Speaker Extraction," *arXiv preprint arXiv:2506.00466*, 2025.

[239] J. Li, K. Zhang, S. Wang, K. A. Lee, M.-W. Mak, and H. Li, "MoMuSE: Momentum Multi-modal Target Speaker Extraction for Real-time Scenarios with Impaired Visual Cues," *arXiv preprint arXiv:2412.08247*, 2024.

[240] B. Gholami, M. El-Khamy, and K.-B. Song, "Knowledge Distillation for Tiny Speech Enhancement with Latent Feature Augmentation," in *Proc. Interspeech*, Kos, Greece, Sept. 2024, pp. 652–656.

[241] A. Moslemi, A. Briskina, Z. Dang, and J. Li, "A Survey on Knowledge Distillation: Recent Advancements," *Machine Learning with Applications*, p. 100 605, 2024.

[242] H.-T. Chiang, K.-H. Hung, S.-W. Fu, H.-C. Kuo, M.-H. Tsai, and Y. Tsao, "Study on the correlation between objective evaluations and subjective speech quality and intelligibility," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Taipei, Taiwan, Dec. 2023, pp. 1–7.

[243] S. Leglaive, M. Fraticelli, H. ElGhazaly, L. Borne, M. Sadeghi, S. Wisdom, M. Pariente, J. R. Hershey, D. Pressnitzer, and J. P. Barker, "Objective and subjective evaluation of speech enhancement methods in the UDASE task of the 7th CHiME challenge," *Computer Speech & Language*, vol. 89, p. 101 685, 2025.

[244] S. Zhang, M. Lei, B. Ma, and L. Xie, "Robust audio-visual speech recognition using bimodal DFSMN with multi-condition training and dropout regularization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May. 2019, pp. 6570–6574.

[245] U. Tiwari, M. Soni, R. Chakraborty, A. Panda, and S. K. Kopparapu, "Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May. 2020, pp. 7194–7198.

[246] Z. Zhao, D. Yang, R. Gu, H. Zhang, and Y. Zou, "Target Confusion in End-to-end Speaker Extraction: Analysis and Approaches," in *Proc. Interspeech*, Incheon, Korea, Sept. 2022, pp. 5333–5337.

[247] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, June. 2021, pp. 920–924.

[248] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[249] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. of Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

LIST OF PUBLICATIONS

The following publications are related to the work in this thesis.

Peer-reviewed Journal Papers

[J1] **R. Sinha**, A-C. Scherer, S. Doclo, C. Rollwage, and J. Rennies, "Evaluation of Speaker-conditioned Target Speaker Extraction Algorithms for Hearing-Impaired Listeners," *Trends in Hearing,* vol. 29, 23312165251365802, 2025.

[J2] **R. Sinha**, C. Rollwage and S. Doclo, "Variants of LSTM cells for single-channel speaker-conditioned target speaker extraction," *EURASIP Journal on Audio, Speech, and Music Processing,* vol. 2024, no. 1, pp. 1-13, 2024.

Peer-reviewed Conference Papers

[C1] **R. Sinha**, C. Rollwage, and S. Doclo, "Real-time Single-channel Speaker-conditioned Target Speaker Extraction using TCN-Conformer with Efficient Self-attention Mechanisms," *in Proc. European Signal Processing Conference (EUSIPCO)*, Palermo, Italy, pp. 1-5, Sept. 2025.

[C2] **R. Sinha**, A-C. Scherer, S. Doclo, C. Rollwage, and J. Rennies, "Subjective performance evaluation of single-channel speaker-conditioned target speaker extraction algorithms for complex acoustic scenes," *in Proc. ITG Conference on Speech Communication*, Aachen, Germany, pp. 1-5, Sept. 2023.

[C3] **R. Sinha**, C. Rollwage, and S. Doclo, "Low-complexity Real-time Single-channel Speech Enhancement Based on Skip-GRUs," *in Proc. ITG Conference on Speech Communication*, Aachen, Germany, pp. 1-5, Sept. 2023.

[C4] **R. Sinha**, M. Tammen, C. Rollwage, and S. Doclo, "Speaker-Conditioning Single-Channel Target Speaker Extraction using Conformer-based Architectures," *in Proc. International Workshop on Acoustic Signal Enhancement (IWAENC),* Bamberg, Germany, pp. 1-5, Sept. 2022.

[C5] **R. Sinha**, M. Tammen, C. Rollwage, and S. Doclo, "Speaker-conditioned target speaker extraction based on customized LSTM cells," *in Proc. ITG Conference on Speech Communication*, VDE, Kiel, Germany, pp. 1-5, Sept-Oct. 2021.