Spatially Selective Speaker Separation Using a DNN With a Location Dependent Feature Extraction

Alexander Bohlender, Ann Spriet, Wouter Tirry, and Nilesh Madhu

***Abstract*—Deep neural networks (DNNs) have proven themselves as an effective means to separate clean speech from noisy mixtures. When there are multiple concurrent talkers, however, unambiguously defining the target output is not trivial, especially if the mixture is single-channel and the talkers are not known in advance. Although this problem can be addressed with permutation invariant training or deep clustering, the performance still suffers in this case. Approaches for compact arrays of multiple microphones can exploit spatial diversity to resolve the ambiguity: a separate output may be generated for each direction of arrival (DOA), or the speaker assignment can be controlled with a location-based training (LBT). Alternatively, we can narrow down the target definition at the input, to perform a spatially selective speaker separation instead of separating all speakers simultaneously. This is achieved by specifying freely adjustable target DOAs. On the one hand, these can be integrated as location-based input features (LBI). On the other hand, the main contribution of this work is a location dependent feature extraction (LDE): we implicitly introduce a DOA dependence in a small part of the DNN by optimizing its parameters for each DOA separately. Experiments demonstrate that LDE outperforms LBT and LBI in terms of instrumental metrics and speech recognition results. A representative audio example is presented for a qualitative impression. An analysis of the spatial selectivity reveals that target and nontarget directions can be distinguished quite well with LDE, which is also verified by recordings of real moving talkers.**

***Index Terms*—spatially selective speaker separation, talker independent, location dependent, ambiguity-free speaker-output assignment, direction of arrival, convolutional recurrent neural network, time-frequency masks.**

I. Introduction

EXTRACTING the clean signal of a target speaker from a mixture with interference and noise is known as speech separation. This is done, e. g., to improve the signal quality in mobile communications, or as a front-end for robust automatic speech recognition (ASR) [1], [2]. A wide variety of deep learning-based methods have been proposed for this task, including approaches designed for single-channel [3], [4] as well as microphone array-based [5], [6] setups. An overview can be found in [7]. When there is only one active talker, the characteristic properties of speech are (implicitly) used to distinguish the target from unwanted signal components. This is achieved by exploiting the availability of representative speech and noise datasets for training.

The problem becomes considerably more challenging when there is an overlap of multiple concurrent talkers. In theory, this *speaker separation problem* can be addressed effectively in the short-time Fourier transform (STFT) domain because of the empirical observation that there is only one dominant speaker in each time-frequency (TF) bin (W-disjoint orthogonality) [8]. For example, we can use a deep neural network (DNN) to obtain masks that locally measure the presence of a speaker in each TF bin in order to attenuate interference, or to directly estimate the complex spectrograms of clean signals [3]. However, since the speakers can be arranged in an arbitrary order, finding a clear (unambiguous) definition of the desired output, which is required for a meaningful training, is not straightforward. Early work on DNN-based speaker separation therefore approached the problem in a talker dependent way [9], where the identities of the target *and* interfering speakers must be known during training, or in a target dependent way [10], where only the interfering speakers are allowed to change.

One established approach for *talker independent* speaker separation is deep clustering (DC) [11]. The underlying idea is to train a DNN to produce embeddings, such that the estimated affinity matrix derived therefrom approximates the ideal binary affinity matrix that indicates in which TF bins the same speaker is dominant. A clustering algorithm like k-means is then applied, of which the output is used to perform the separation. Deep attractor networks [12], which also generate embeddings, have further improved the performance compared to the original DC. In contrast, with (utterance-level) permutation invariant training (PIT) [13], [14], the ambiguous assignment of the speakers to the output channels (permutation ambiguity problem) is addressed by considering all possible permutations during training, and selecting the one that yields the lowest error. At test time, one output is then obtained for each speaker directly. It is reported in [14] that deep attractor networks and utterance-level PIT perform comparably. The differences between the two approaches are therefore mainly conceptual: PIT is simpler during inference as no clustering or attractor point estimation is needed, whereas methods based on latent representations do not require defining a (maximum) number of separable speakers during training.

Here, we focus on mixtures captured by a compact array of microphones. Naturally, both DC and PIT can be extended to such a *multichannel* setup [15], [16]. However, these extensions only make indirect use of spatial information to address the permutation ambiguity. As speakers are typically not co-located in space, a powerful alternative can be to find a unique speaker-to-output assignment based on spatial diversity.

The approach of [17] takes advantage of the close connection between narrowband (frequency bin-wise) direction

A. Bohlender, and N. Madhu are with IDLab, Department of Electronics and Information Systems, Ghent University - imec, 9000 Ghent, Belgium (e-mail: alexander.bohlender@ugent.be; nilesh.madhu@ugent.be).

A. Spriet, and W. Tirry are with Goodix Technology (Belgium) B.V., 3000 Leuven, Belgium (e-mail: aspriet@goodix.com; wtirry@goodix.com).

of arrival (DOA) estimation and source separation. The space around the array is partitioned into a discrete angular grid, and probabilities of speech activity for each of these angles and for all TF bins are estimated with a U-Net. These probabilities then serve as TF masks, which can be used to extract a speaker from a selected discrete direction. Similar to that, our approach of [18] also associates each output channel with one DOA. Whereas [17] performs DOA estimation and speaker separation jointly, [18] relies on prior knowledge of the DOAs to select the correct output channel for each speaker. However, both approaches have in common that a DNN is trained to generate a separate output for *all* directions, even those without source activity. This implies a high redundancy, which limits the achievable performance due to the suboptimal use of the computational power of the DNN. Recently, a location-based training (LBT) [19], [20] was proposed, where the output channel assignment is performed in an increasing order of the speakers' azimuth angles of arrival *or* their distances from the array. Thus, with just one output channel for each speaker, the approach can more effectively focus on the signals of interest than [17], [18]. LBT was also shown to outperform PIT consistently, which suggests that a clearly defined output channel assignment is still beneficial.

Based on this observation regarding the usefulness of an ambiguity-free *output* representation, we hypothesize that already defining the target more clearly at the *input* would further enable the DNN to focus on the most relevant information. With the availability of multiple microphones, we can resort to the source DOAs to define one (or more) target speakers, and consider all other speakers as interferers. We term this the (spatially) *selective speaker separation (SSS)* problem. If the DOAs of interest are not known in advance, they can be estimated before the separation, or both can be combined in an end-to-end system [6]. The designated target DOAs can be integrated into the DNN, e. g., in the form of additional location-based input features (LBI). This is done in [21], where the target DOAs are encoded by the corresponding ideal interchannel phase differences. A very similar so-called angle feature is often used as well [6], [22], [23]. In the end-to-end architecture of [6], the need for prior information on the speaker locations is avoided by letting an attention module select the location-based input features.

To address the potential lack of optimality of *hand-crafted* representations, the DNN could also learn to convert the DOA information to adequate input features on its own [24], [25]. It is demonstrated in [24] that such a DNN-designed feature representation can improve the performance compared to the (hand-crafted) phase difference-based location features. To jointly consider DOA information (regardless of the encoding) and microphone signals, straightforwardly, both can be concatenated directly at the input. Alternatively, [25] recently proposed to use the DOA-based features to determine the initial state of long short-term memory (LSTM) layers within the DNN architecture.

In this work, we formulate the SSS problem such that a target DOA *range* can be selected dynamically, where the number of speakers contained therein is arbitrary. A possible application could be, e. g., the extraction of (a mixture of) all speakers within the area covered by a camera, so as to obtain matched audio and video. Individual speakers can still be extracted as well, if a (narrow) range of target DOAs is chosen based on the output of a separate DOA estimation step. The main contribution of this paper is a *location dependent feature extraction (LDE)*, an alternative to LBI where we propose to use the DOAs to determine *how* the DNN processes the microphone signals. This is achieved by optimizing a subset of the DNN parameters for each direction separately. Specifically, we choose to introduce a DOA dependence only in the *first* layer of the considered convolutional recurrent neural network (CRNN). This way, we can limit the impact of the location dependent feature extraction on the complexity of the network while ensuring the target is clearly defined from the beginning.

Our experiments verify that the proposed LDE improves the performance of an *otherwise unchanged system* compared to LBT and LBI in terms of speech quality, intelligibility, and ASR results. Spectrograms and audio files for a representative example allow for a qualitative appreciation of the differences between the methods, and demonstrate that the detailed spectro-temporal structure of the target is better captured with LDE, thereby reducing speech distortion. We also analyze the spatial selectivity that the system gains through LDE, in order to provide insight into the appropriate choice of the target DOA range, and to understand the limitations in challenging conditions.

In Section II, the general source separation problem and the CRNN-based system used in this work are introduced. Section III reviews the state-of-the-art for separating all speakers simultaneously. Then, in Section IV, suitable approaches for the location-based formulation of the SSS are discussed, which require only one output channel. Our training paradigm, that is essential for the DNN to learn how to perform the LDE, is outlined in Section V. An experimental analysis based on in-house recordings is conducted in Section VI. We also provide results for the LibriCSS dataset of [26] in Section VII. Section VIII concludes the paper.

II. Neural Network-Based Source Separation

A. Problem Formulation

We consider a compact array of N microphones that capture sound from J localized sources and background noise. The focus of this work is on speech, in which case the sources correspond to different talkers. Each source contribution consists of a direct-path component and a reverberation component that are denoted by $S'_{j,n}(f,t)$ and $S''_{j,n}(f,t)$, respectively, in the STFT domain. We use f to denote the discrete frequency index for a discrete Fourier transform (DFT) size of F, t to denote the frame index, j to denote the source index, and n to denote the microphone index. Consequently, with the additive noise $V_n(f,t)$, the signal model for the nth microphone is given by

$$Y_n(f,t) = \sum_{1\leq j\leq J} S'_{j,n}(f,t) + S''_{j,n}(f,t) \;+\; V_n(f,t). \qquad (1)$$

From this mixture, we aim to extract one signal containing all the direct-path components $S'_{j,n}(f,t)$ with $j \in \mathcal{J}(t)$, where $\mathcal{J}(t) \subseteq \{1,2,\ldots,J\}$ is a defined subset of the sources. The

remaining $j \notin \mathcal{J}(t)$ are interferers, which should be suppressed along with reverberation and noise. Thus, we define

$$S_n(f,t) = \sum_{j \in \mathcal{J}(t)} \gamma_j S'_{j,n}(f,t) \tag{2}$$

to be the target signal, and the unwanted components are

$$X_n(f,t) = \sum_{j \notin \mathcal{J}(t)} S'_{j,n}(f,t) + \sum_{1 \le j \le J} S''_{j,n}(f,t) + V_n(f,t). \tag{3}$$

In (2), γ_j is a time-invariant scaling factor set to

$$\gamma_j = \sqrt{\frac{\sum_{f,t,n} \left| S'_{j,n}(f,t) + S''_{j,n}(f,t) \right|^2}{\sum_{f,t,n} \left| S'_{j,n}(f,t) \right|^2}}. \tag{4}$$

Its purpose is to ensure that the dynamic range of the desired output is consistent across different direct-to-reverberant ratios [18]. This can help prevent the DNN from completely suppressing distant sources where the energy of the direct path may be relatively small compared to the reverberation component. Note that the target signal (2) is a mixture of multiple sources when the set of target source indices has a cardinality $|\mathcal{J}(t)| > 1$. In Sections III and IV, we will use different formulations of the speaker separation problem, which lead to different definitions of $\mathcal{J}(t)$.

We now define a reference channel, and omit the microphone index where applicable. Often, an arbitrary microphone is used for this purpose. However, this could cause a bias in favor of the selected microphone, and make it impossible to reconstruct the target with a mask-based approach in TF bins where the reference microphone amplitude is close to 0 due to nulls in the acoustic transfer function. Here, the root mean square of the N channels is therefore taken for the *magnitude* instead. We define

$$S(f,t) = \sqrt{\frac{1}{N} \sum_{1 \le n \le N} |S_n(f,t)|^2}\, \mathrm{e}^{\jmath \angle S_1(f,t)} \tag{5a}$$

$$X(f,t) = \sqrt{\frac{1}{N} \sum_{1 \le n \le N} |X_n(f,t)|^2}\, \mathrm{e}^{\jmath \angle X_1(f,t)} \tag{5b}$$

$$Y(f,t) = \sqrt{\frac{1}{N} \sum_{1 \le n \le N} |Y_n(f,t)|^2}\, \mathrm{e}^{\jmath \angle Y_1(f,t)}. \tag{5c}$$

The reference *phase* still comes from one microphone ($n = 1$).

B. Time-Frequency Masking

Various methods to estimate the target signal $\widehat{S}(f,t)$ based on the output of a DNN have been discussed in the literature [3], [7]. Whereas some only enhance the magnitude but retain the noisy phase, for an improved speech quality clean magnitude and phase can be estimated jointly. Because we are mainly interested in comparing methods to cope with the permutation ambiguity in this work, for simplicity only ideal ratio masks (IRMs) are estimated here. Note that this is done without loss of generality: with appropriately chosen input and desired output, e. g., a complex spectral mapping (CSM) [3], [27] could easily be performed instead.

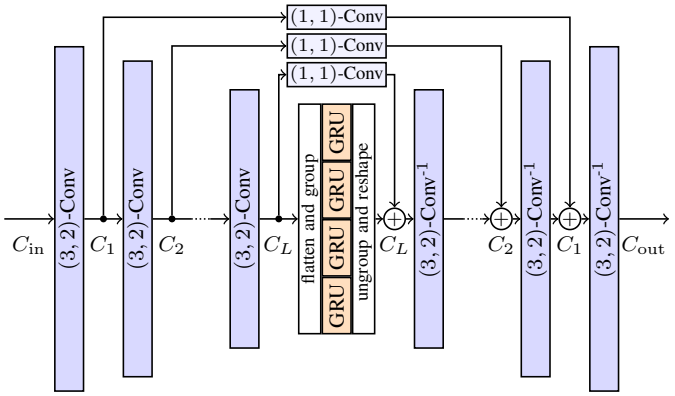

Fig. 1. CRUSE [4] consists of L encoder-decoder modules with additive skip connections, and parallel recurrent layers at the network bottleneck. Typically, the number of feature maps (channels) $\{C_{\text{in}}, C_1, \ldots, C_L\}$ increases gradually throughout the encoder, while the frequency dimension is compressed.

The IRM which we use as training target is given by

$$\mathcal{M}(f,t) = \left[\frac{|S(f,t)|^2}{|S(f,t)|^2 + |X(f,t)|^2} \right]^{\beta}, \tag{6}$$

such that $\mathcal{M}(f,t) \in [0,1]$. We choose $\beta = 1$ for a strong suppression, at the cost of an increased target distortion compared to the often used $\beta = 0.5$. The output signal is obtained by applying the estimated mask $\widehat{\mathcal{M}}(f,t)$ as a gain:

$$\widehat{S}(f,t) = \widehat{\mathcal{M}}(f,t)\, Y(f,t). \tag{7}$$

C. Convolutional Recurrent U-Net for Speech Enhancement

Fig. 1 shows the convolutional recurrent U-Net for speech enhancement (CRUSE), which was proposed in [4] as a computationally efficient version of the encoder-decoder architecture of [28]. Input and output are tensors with a size of $F' \times T \times C_{\text{in}}$ and $F' \times T \times C_{\text{out}}$, respectively, where $F' = F/2 + 1$ is the number of frequency bins up to the Nyquist frequency, and T is the (arbitrary) number of frames. The choice of C_{in} and C_{out} is discussed below.

The encoder consists of L convolutional layers that operate over the frequency and time dimensions, where a leaky ReLU activation function is applied after each. The size of the *causal* filters is set to 2 for the time dimension, and 3 for the frequency dimension. To compress the frequency axis, a stride of 2 is used for this dimension only, whereas the number of frames remains unchanged. The number of channels is increased gradually throughout the encoder layers. In [4], $C_1 = 16$ after the first layer, followed by repeated doubling ($C_l = 2\,C_{l-1}$) up to a specified maximum of $C_{\max}$.

A recurrent layer between the encoder and the decoder incorporates long-term temporal context information. To improve computational efficiency, in this layer only, the features from all subbands are divided into multiple groups, which are processed in parallel by gated recurrent units (GRUs). Experiments conducted in [4] indicate that the performance does not deteriorate significantly with up to 4 parallel GRUs.

The decoder performs operations that are inverse to the encoder so as to create a symmetric architecture. As Fig. 1 shows, skip connections between encoder and decoder bypass the

inner layers of the network. These preserve local information, which is especially beneficial when enhancing speech in the STFT domain due to the importance of accurately capturing the detailed TF structure of the target speech for the *perceived* audio quality [18]. If a learnable channel-wise scaling and bias, which is implemented as a grouped convolution with a $(1,1)$-kernel, is incorporated within, the performance achieved with *additive* skip connections is comparable to that of the more complex concatenation along the feature dimension [4].

Application to a multichannel setup: In [4], where CRUSE is used for single-channel speech enhancement, the log-power spectrogram is the only input ($C_{\text{in}} = 1$). In contrast, we have N microphone signals $\mathbf{Y}(f,t) = [Y_1(f,t), \ldots, Y_N(f,t)]^T$ in this work, which can be decomposed into two factors

$$\mathbf{Y}(f,t) = \underbrace{\frac{\mathbf{Y}(f,t)}{\|\mathbf{Y}(f,t)\|_{\ell_2}}}_{\text{spatial}} \underbrace{\|\mathbf{Y}(f,t)\|_{\ell_2}}_{\text{spectral}} . \tag{8}$$

The first factor, which is referred to as the *directional statistics* in [29], captures the phase and level differences between the microphones, and thus the *spatial* information. A variant thereof is used as DNN input, e. g., in [30], where a whitening is performed in addition. The second factor is the reference channel magnitude, which can serve as the *spectral* information. The magnitude range may be compressed by taking the logarithm, whereas real and imaginary part of the directional statistics are inherently confined to $[-1,1]$. To facilitate the exploitation of all available information, we use the decomposition of (8) to define the input feature vector

$$\mathbf{Z}(f,t) = \begin{bmatrix} \Re\{\mathbf{Y}(f,t)\} / \|\mathbf{Y}(f,t)\|_{\ell_2} \\ \Im\{\mathbf{Y}(f,t)\} / \|\mathbf{Y}(f,t)\|_{\ell_2} \\ \log \|\mathbf{Y}(f,t)\|_{\ell_2} - E\left\{\log \|\mathbf{Y}(f,t)\|_{\ell_2}\right\} \end{bmatrix} \tag{9}$$

of length $C_{\text{in}} = 2N + 1$. The expected value of the log-magnitude, which is approximated by

$$E\left\{\log \|\mathbf{Y}(f,t)\|_{\ell_2}\right\} = \frac{1}{\overline{T}F'} \sum_{\substack{0 \le f' < F' \\ t-\overline{T} < t' \le t}} \log \|\mathbf{Y}(f',t')\|_{\ell_2} \tag{10}$$

during both training and inference, is subtracted in order to guarantee scale invariance. The parameter $\overline{T}$ is chosen such that we obtain a moving average length of $0.3\,\text{s}$.

Empirically, we observe that $C_1 = 16$ channels after the first encoder layer significantly limit the ability of the network to make full use of the information contained in the provided $2N + 1$ input features. Therefore, we choose $C_1 = 64$ instead. Two different configurations are considered in the following: a low-complexity (LC) setup ($C_{\max} = 64$, $L = 4$), and a high-complexity (HC) alternative ($C_{\max} = 256$, $L = 5$).

III. Established Multi-Output Speaker Separation

When there is just one speaker, only a single output channel is required ($C_{\text{out}} = 1$). For the problem of separating *multiple* speakers, however, finding an unambiguous interpretation of the output channels is not straightforward. Several approaches proposed in prior work are discussed in this section.

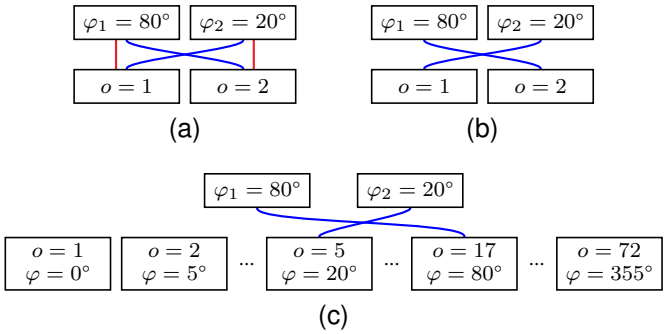

Fig. 2. Speaker-to-output assignment during training for an example with $J = 2$ speakers. (a) Permutation invariant training (PIT): consider all possible permutations (here: represented by different colors). (b) Location-based training (LBT): select permutation that sorts the azimuth angles. (c) Location-based output channels (LBO): each output corresponds to a fixed angle.

A. Permutation Invariant Training (PIT)

In principle, one mask can be generated for each speaker directly with an appropriate number of output channels ($C_{\text{out}} = J$). The target mask $\mathcal{M}^{(o)}(f,t)$ for output $o \in \{1, 2, \ldots, J\}$ is obtained by setting $\mathcal{J}^{(o)}(t) = \{o\}$ in the definition of the target signal (2).

However, the speaker-output assignment is arbitrary, in general. For example, it is irrelevant whether the first output channel generates a mask for speaker $j = 2$ and the second output channel for speaker $j = 1$, or vice versa. When the DNN has no means to determine which speaker corresponds to which output channel, it is not sensible to assume a fixed channel assignment during training. Instead, a permutation invariant training (PIT) can be performed [13], [14], for which the loss function is given by

$$\mathcal{L}_{\text{PIT}} = \min_{P \in \mathcal{S}_J} \sum_{1 \le o \le J} \mathcal{E}\left(\mathcal{M}^{(o)}(f,t), \widehat{\mathcal{M}}^{(P(o))}(f,t)\right), \tag{11}$$

where $\mathcal{S}_J$ is the symmetric group of degree J, i. e., the set of all $J!$ permutations of $\{1, 2, \ldots, J\}$, and $\mathcal{E}(\cdot)$ is a suitable error function, e. g., the mean square error (MSE). This is illustrated in Fig. 2a. Whereas the original PIT [13] only required a consistent permutation within a fixed number of successive frames (a so-called meta-frame), recurrent neural networks (RNNs) were later used for an extension to signals of arbitrary length. Due to the use of whole utterances for training, [14] refers to this as utterance-level PIT (uPIT).

B. Location-Based Training (LBT)

The motivation for PIT, which was originally proposed for a single-microphone setup, is the difficulty of unambiguously assigning the speakers to the output channels. In microphone array-based source separation, spatial information can be leveraged to address this problem. A location-based training (LBT) was proposed in [19], where the sources are assigned to the output channels in an increasing order of their azimuth angles of arrival *or* their distances from the array. Further experiments conducted in [20] show that azimuth-based training outperforms distance-based training, except when a too small difference between the azimuth angles makes it difficult to use

this criterion to distinguish between two sources. It was found that a fusion of azimuth- and distance-based training enables a further improvement. Regardless of the selected criterion, [20] showed that LBT outperforms PIT consistently, which leads us to select LBT as a baseline in the following.

In this work, we focus on *azimuth-based* criteria. For comparability, this is done consistently for LBT and other location-based approaches. Therefore, we again have $C_{\text{out}} = J$, but arrange the output channels according to the unique azimuth angles of arrival, which are denoted by $\varphi_1(t), \ldots, \varphi_J(t)$ with $0° \leq \varphi_j(t) < 360°$ for all j. We can then formally define the target source indices for the J output channels as

$$\mathcal{J}^{(o)}(t) = \{P_\varphi^t(o)\}, \tag{12}$$

where the permutation P_φ^t sorts the azimuth angles at time t:

$$\varphi_{P_\varphi^t(1)}(t) < \varphi_{P_\varphi^t(2)}(t) < \ldots < \varphi_{P_\varphi^t(J)}(t).$$

The LBT loss is then simply given by

$$\mathcal{L}_{\text{LBT}} = \sum_{1 \leq o \leq J} \mathcal{E}\left(\mathcal{M}^{(o)}(f,t), \widehat{\mathcal{M}}^{(o)}(f,t)\right). \tag{13}$$

This is illustrated in Fig. 2b.

A shortcoming of this approach is that sudden changes in the speaker-to-output assignment can occur, e. g., when a new talker starts to speak after the initial channel allocation, and their azimuth angle is smaller than that of an already active talker. It should also be noted that in the case of both PIT and LBT, the number of separable sources is limited by the number of output channels.

C. Location-Based Output Channels (LBO)

In [18], we proposed to associate each output channel with one DOA rather than one speaker. This is referred to as location-based output channels (LBO) here. A discrete grid

$$\Psi = \{\psi_1, \ldots, \psi_D\} \tag{14}$$

is defined for the DOAs, and $C_{\text{out}} = D$. For example, when the array is planar and only the azimuth angle is considered with a resolution of 5°, we obtain $D = 72$ possible DOAs:

$$\Psi = \{0°, 5°, \ldots, 355°\}. \tag{15}$$

When $\phi_j(t) \in \Psi$ is the *discretized* DOA of the jth source (for the grid of (15): obtained by quantizing the azimuth angle $\varphi_j(t)$), the sources are assigned to the output channels with

$$\mathcal{J}^{(o)}(t) = \{j : \phi_j(t) = \psi_o\}, \tag{16}$$

and the loss is given by

$$\mathcal{L}_{\text{LBO}} = \sum_{1 \leq o \leq D} \mathcal{E}\left(a^{(o)}(t)\mathcal{M}^{(o)}(f,t), a^{(o)}(t)\widehat{\mathcal{M}}^{(o)}(f,t)\right), \tag{17}$$

where

$$a^{(o)}(t) = \begin{cases} 1, & \mathcal{J}^{(o)}(t) \neq \emptyset \\ 0, & \text{else.} \end{cases} \tag{18}$$

This is shown in Fig. 2c. A weighting $a^{(o)}(t) = 0$ is used to exclude directions with no active speaker, for which $\mathcal{J}^{(o)}(t) = \emptyset$, and thus $\mathcal{M}^{(o)}(f,t) = 0$, i. e., these output channels are viewed as *don't cares* in the training. This is necessary since explicitly minimizing the output for inactive directions would introduce a strong bias in favor of small mask values.

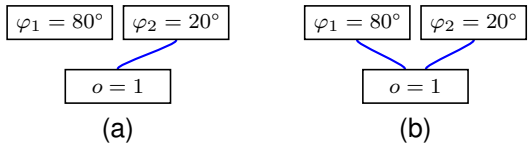

Fig. 3. Selective speaker separation (SSS): a source is considered to be part of the target output if and only if $\phi_j \in \Psi^*$, where ϕ_j is the corresponding *discrete* source DOA (e. g., $\varphi_j = 59° \Rightarrow \phi_j = 60°$ for the discrete grid of (15)). (a) $\Psi^* = \{10°, 15°, 20°\}$ (b) $\Psi^* = \{15°, 20°, 25°, 75°, 80°, 85°\}$

The LBO approach has two major drawbacks. First, it requires the knowledge of the source DOAs for the output channel selection, but this knowledge is not actually exploited within the mask estimation. Secondly, the generation of D masks, out of which only $J \ll D$ are needed to extract estimates of the source signals, is highly redundant.

IV. Spatially Selective Speaker Separation

The methods discussed in Section III have in common that the number of output channels specified during training does not always match the number of sources to be separated. This can lead to an inefficient use of the DNN (too many outputs), or make it impossible to separate all sources (too few outputs). Here, we propose a more focused problem formulation: our aim is to generate a single output containing a subset of the sources, which can be *chosen dynamically* during inference. We refer to this as selective speaker separation (SSS). To obtain, again, a location-based criterion, a set of target DOAs $\Psi^*(t) \subseteq \Psi$ is defined, where Ψ is the discrete grid of (14). Only one output channel ($C_{\text{out}} = 1$) is then needed, for which

$$\mathcal{J}(t) = \{j : \phi_j(t) \in \Psi^*(t)\}. \tag{19}$$

are the target source indices. Thus, no output is generated for unwanted sources or directions. When there is more than one source with a DOA $\phi_j(t) \in \Psi^*(t)$, the desired output is a *mixture* of all of these sources. Individual sources can still be extracted by appropriately selecting $\Psi^*(t)$ based on the output of a separate DOA estimation step. In contrast to LBO, we integrate $\Psi^*(t)$ into the DNN to exploit this information within the mask estimation. In addition, the trade-off between spatial selectivity and robustness to DOA estimation errors can be dynamically controlled by widening or narrowing the range of target DOAs. The described target definition can also be useful for other applications, e. g., automotive: focus on the driver and the front seat passenger of a car, while suppressing sound from the back. The computation of the loss

$$\mathcal{L}_{\text{SSS}} = \mathcal{E}\left(a(t)\mathcal{M}(f,t), a(t)\widehat{\mathcal{M}}(f,t)\right) \tag{20}$$

with

$$a(t) = \begin{cases} 1, & \mathcal{J}(t) \neq \emptyset \\ 0, & \text{else} \end{cases} \tag{21}$$

analogously to (18) is illustrated in Fig. 3.

A. Location-Based Input Features (LBI)

To generate an output that is dependent on the specified target DOAs, $\Psi^*(t)$ must be integrated into the DNN. This can be straightforwardly achieved with additional input features that are concatenated with $\mathbf{Z}(f,t)$. Several approaches have been proposed to obtain a suitable TF representation of the source DOAs, either to exploit their knowledge for an improved separation, or to select one particular target source. In [21], the *measured* (microphone signal) interchannel phase differences are subtracted from the *ideal* interchannel phase differences for a plane wave incident from the specified DOA. The cosine thereof is taken to address the discontinuity of phase due to the 2π-periodicity. This is equivalent to the cosine similarity between the steering vector for the selected DOA and $\mathbf{Y}(f,t)/Y_1(f,t)$, which is referred to as the *angle feature* in [22]. It is worth noting that the angle feature can also be used to support an approach with LBO. In [23], a multi-look enhancement network (MLENet) was proposed where the input includes angle features for several look directions, for each of which an estimate of the nearest speaker is returned.

The disadvantage of a hand-crafted TF representation of the DOAs is that optimality is not ensured. Therefore, [24] proposed to use a multi-hot vector $\mathbf{Z}_\phi(t) \in \{0,1\}^D$ as input, where the dth element is given by

$$\Big[\mathbf{Z}_\phi(t)\Big]_d = \begin{cases} 1, & \exists j : \phi_j(t) = \psi_d \\ 0, & \text{else,} \end{cases} \tag{22}$$

and to then let the DNN learn a suitable TF representation on its own. For this purpose, an additional fully connected (FC) layer is added before CRUSE. It is shown in [24] that this can improve performance compared to the hand-crafted representation based on the ideal phase differences. A similar approach was adopted by [25], where FC layers process a one-hot vector specifying a single target DOA in order to determine the initial state of the two LSTM layers that the employed DNN consists of. Thus, the *core* idea of [24] and [25] is the same: leave it to the DNN to design appropriate features based on DOA information. However, the approach of [25] is specific to LSTM (or other type of RNN) architectures, and a time variance in the target DOA can only be realized by overwriting the LSTM state, whereby potentially useful information is discarded. In the following, we therefore consider the generally applicable concatenation of DOA-based and other inputs to represent this class of approaches.

Here, for the SSS, the aim is to encode information on an arbitrary number of target DOAs (rather than a fixed number of source DOAs), which can be achieved with the multi-hot vector representation. Similar to (22), we therefore set

$$\Big[\mathbf{Z}^*(t)\Big]_d = \begin{cases} 1, & \psi_d \in \Psi^*(t) \\ 0, & \text{else.} \end{cases} \tag{23}$$

This is referred to as location-based input features (LBI) in the following. The combination of the microphone signal-based features $\mathbf{Z}(f,t)$ and those determined based on the specified target DOAs $\mathbf{Z}^*(t)$ is visualized in Fig. 4. An FC layer generates C^* elements *per frequency* from the vector $\mathbf{Z}^*(t)$.

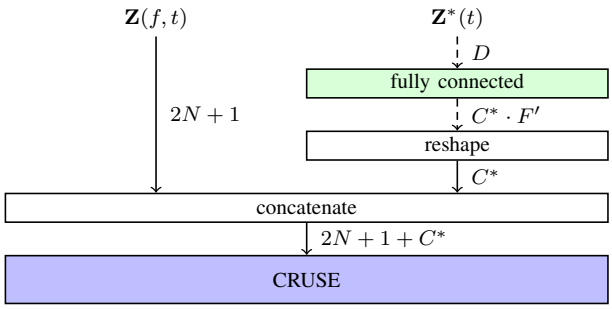

Fig. 4. The multi-hot input vector $\mathbf{Z}^*(t)$ is converted to a time-frequency representation with an FC layer. The total number of input features increases to $2N+1+C^*$, where C^* is a hyperparameter. Dashed lines represent tensors with two dimensions (no frequency), and solid lines for three dimensions.

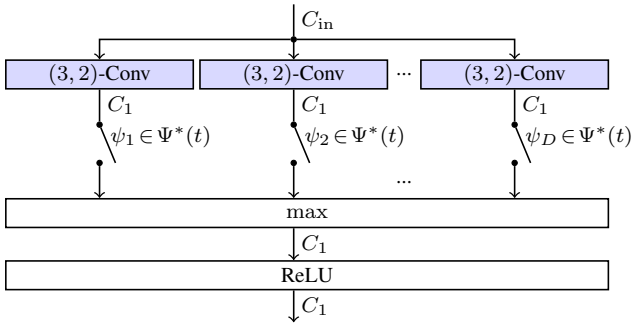

Fig. 5. Proposed location dependent feature extraction (LDE) in the first encoder layer: separate optimization of the learnable parameters for each DOA. The maximum of all directions out of the specified target DOA range $\psi_d \in \Psi^*(t)$ is taken for each of the C_1 generated features. The CRUSE architecture remains unchanged otherwise (see Fig. 1).

Empirically, we set $C^* = 12$ for the LC setup ($C_{\max} = 64$), and $C^* = 18$ for the HC setup ($C_{\max} = 256$).

B. Location Dependent Feature Extraction (LDE)

Microphone signal spectra and DOAs are different types of information. Therefore, it may not be ideal to combine both directly. On the other hand, it would limit the ability of the DNN to effectively focus on the signal of interest if the target DOAs were only integrated deeper in the network architecture.

As an alternative, we propose a novel location dependent feature extraction (LDE), where a small subset of the learnable parameters is optimized separately for each DOA. Here, we choose to incorporate DOA dependencies within the *first encoder layer* for two reasons: First, the immediate knowledge of the target DOAs allows the DNN to extract only relevant information, thereby avoiding the implicit need for a holistic description of the acoustic scene (including unwanted speakers and directions). Secondly, when $C_{\text{in}} = 2N+1$ and $C_1 = 64$, the input size of this layer is quite small compared to the rest of the network (assuming an array of $N \ll C_1$ microphones), so that the cost of introducing a DOA dependence here is low.

The location dependent first convolutional layer is realized as depicted in Fig. 5. For each direction $\psi_d \in \Psi^*(t)$, we obtain C_1 features. The DOA dimension created in the process is condensed with a pooling operation. We choose to take the *maximum* here to allow sources over a wider spatial region to

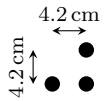

Fig. 6. An array geometry with $N = 3$ microphones is considered.

be retained as more different directions are included in $\Psi^*(t)$. At the same time, incorporating additional directions with no speech activity should have a limited effect on the output, which can be more easily achieved with a maximum operation as opposed to, e. g., summation.

Compared to the original CRUSE [4], the LDE increases the number of trainable parameters of the first encoder layer by a factor of D. The number of multiply-accumulate (MAC) operations per frame only increases by a factor of $|\Psi^*(t)| \leq D$. No changes are required with respect to the remainder of the network of Fig. 1 (following the first encoder layer).

V. Training Setup

We use simulated microphone signals as training data. In the simulation, the $N = 3$ microphones of the triangular array geometry shown in Fig. 6 are placed at a fixed height above the ground. Note that any array configuration can be used, provided it is the same during training and inference.

The generated additive mixtures of $J \in \{1, 2\}$ speakers and noise follow the signal model of (1). For the clean speech, the TIMIT [31] and PTDB-TUG [32] datasets are used. After removing silent segments, if necessary, multiple utterances are concatenated. The resulting source signals are $2\,\mathrm{s}$ long (sampling rate $f_s = 16\,\mathrm{kHz}$).

We consider only the azimuth angles, and use (15) as the DOA grid. Room impulse responses (RIRs) are simulated with [33] for 10 rooms (reverberation: between $T_{60} = 0.2\,\mathrm{s}$ and $0.8\,\mathrm{s}$), 7 array positions within each room, and 4 source-array distances for each of the $D = 72$ DOAs. A smaller (unique) set of RIRs is simulated for the validation. The source signals are convolved with the RIRs to get reverberant signals. One RIR is randomly selected for each source (same room, but different DOAs and distances). More details regarding the room simulation parameters can be found in [34]. The direct-path components needed to compute the target masks are determined by convolving with the respective anechoic RIRs.

For the STFT, we use an $F = 512$-point DFT ($F' = 257$). The frame length and frame shift are set to 512 samples ($32\,\mathrm{ms}$) and 160 samples ($10\,\mathrm{ms}$), respectively. A square-root Hann window function is employed in the analysis.

The approach of [35] is used to simulate a spherically isotropic noise field. The generated spectrally white and spatially diffuse noise is mixed with the speech signals at a random sources-to-noise ratio (SNR), i. e., the ratio of the energies of all sources (including reverberation) to additive noise, taken from a uniform distribution $\mathcal{U}(0, 30)\,\mathrm{dB}$.

To remove unwanted signal components effectively, the CRUSE output is interpreted as the log-masks: $\log \widehat{\mathcal{M}}(f, t)$ for the SSS, or $\log \widehat{\mathcal{M}}^{(o)}(f, t)$ when there are multiple output channels. Both the estimated and the target log-masks are clipped at $\log(0.01)$ (lower bound) and $\log(1)$ (upper bound). The loss is then computed with the MSE error function

$$\mathcal{E}\left(\mathcal{M}(f,t), \widehat{\mathcal{M}}(f,t)\right) = \sum_{f,t} \left(\log \mathcal{M}(f,t) - \log \widehat{\mathcal{M}}(f,t)\right)^2. \tag{24}$$

Like [4], we use an AdamW optimizer [36] with learning rate $8 \cdot 10^{-5}$ and weight decay 0.1 for training. Batch normalization is applied, where each batch comprises 5 signals of 200 frames. The model parameters are saved regularly during training, so that the snapshot with the lowest validation loss can be selected in the end.

A. Choice of the Target DOAs

For the DNN to learn the relation between the specified target DOAs and the corresponding expected output, it plays an important role how $\Psi^*(t)$ is chosen during training. This choice can be made with the intended application in mind. If the aim is to extract a single source with a known DOA, $\Psi^*(t)$ could be selected such that only this one direction is included. To address the more general case where we are interested in one or more sources in a freely selectable spatial region, a more generic approach is needed. Here, we define a contiguous range of target DOAs based on two parameters: a center direction φ_c and a tolerated deviation $\Delta\varphi$, i. e.,

$$\Psi^*(t) = \{\varphi \in \Psi : |\mathrm{angdif}(\varphi, \varphi_\mathrm{c})| \leq \Delta\varphi\}, \tag{25}$$

with the difference between two azimuth angles given by

$$\mathrm{angdif}(\varphi', \varphi'') = \mathrm{mod}(\varphi' - \varphi'' + 180°, 360°) - 180°, \tag{26}$$

where $\mathrm{mod}(\cdot)$ is applied to reflect that the azimuth angle is uniquely defined only on a 360°-range. The target (25) may be compared to a beamformer, where φ_c is the steering direction and $2\Delta\varphi$ is the (frequency-invariant) beamwidth. With a probability of 50%, φ_c is set to either of the true source DOAs ($\varphi_\mathrm{c} \in \{\varphi_1(t = 0), \ldots, \varphi_J(t = 0)\}$). Thus, scenarios with a source in the center of the specified target range are strongly represented. With the remaining 50% probability, φ_c is selected randomly in Ψ. In practice, the target is typically formed by a limited range of angles ($\Delta\varphi \ll D$), at least when extracting individual speakers. In the training, we therefore set

$$\Delta\varphi = \lfloor G - 1 \rfloor \cdot 5°, \tag{27}$$

where $G \sim \mathcal{G}(1, D/4 + 2)$ follows a log-uniform distribution with support $[1, D/4 + 2]$. The resulting cumulative distribution function (CDF) for $\Delta\varphi$ is visualized in Fig. 7.

Note that the source and target DOAs are *fixed* in the simulation, i. e., the training setup does not account for moving sources, and $\Psi^*(t)$ is the same for all t. Our experiments indicate that this is a valid simplification since we observe a good generalization to time-variant source and target DOAs (also when, e. g., the role of target and interferer suddenly reverses). This will also be demonstrated in Section VI. In the case of LBT, we observe that the output channel assignment changes rapidly depending on the current DOAs.

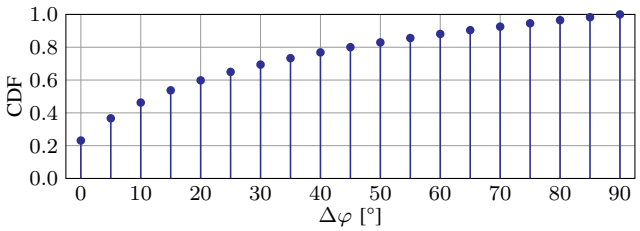

Fig. 7. CDF for the tolerated deviation $\Delta\varphi$ from the target direction in the training data generation. For an effective extraction of sources from a narrow angular range, small values are chosen more commonly.

TABLE I
NUMBER OF TRAINABLE PARAMETERS AND MACS PER FRAME. THE SSS REQUIRES MULTIPLE RUNS TO EXTRACT ALL INDIVIDUAL SOURCES.

		$C_{\max}$	L	parameters	MACs per frame to extract 1 source	MACs per frame to extract 2 sources
LC-	LBT	64	4	1.5 M	7.2 M	7.2 M
HC-	LBT	256	5	6.9 M	41.0 M	41.0 M
LC-	LBI	64	4	1.8 M	7.8 M	15.5 M
HC-	LBI	256	5	7.2 M	41.8 M	83.7 M
LC-	LDE	64	4	1.7 M	8.5 M	17.0 M
HC-	LDE	256	5	7.1 M	42.3 M	84.6 M

VI. EVALUATION

Experimental evaluation in Sections VI-B to VI-D is carried out based on microphone signals generated using RIRs and noise recorded in a real room, unlike the shoebox room acoustics of the training setup. In this manner, we can come close to a real recording while still having access to the clean reference signals. The setup is outlined in Section VI-A. In addition, in Section VI-E, we further validate our findings with live recordings of moving talkers.

A. Experimental Setup

RIRs were recorded in a meeting room with approximate dimensions $7.50\,\mathrm{m} \times 5.00\,\mathrm{m} \times 2.65\,\mathrm{m}$ and a reverberation time of about $0.66\,\mathrm{s}$. Loudspeaker and microphone array were set up such that we have azimuth angles $\varphi \in \{0°, 20°, \ldots, 180°\}$ for source-array distances of $1\,\mathrm{m}$ and $2\,\mathrm{m}$, and $\varphi \in \{40°, 60°, \ldots, 140°\}$ for a distance of $3\,\mathrm{m}$. Relatively diffuse noise was recorded in a lecture room, where multiple loudspeakers simultaneously played back the pub noise signal from the ETSI background noise database [37]. We also recorded ambient noise under quiet conditions.

As our formulation of the location-based SSS requires only one output channel corresponding to a defined spatial region, the number of target and interfering speakers can be arbitrary, in general. For conciseness, we nevertheless focus on scenarios with exactly $J = 2$ concurrent speakers. From the TSP speech database [38], 5 utterances of the same talker are concatenated to obtain one source signal. We consider a setup with time-invariant and a setup with temporally varying source DOAs. For the latter, the azimuth angle of a source is changed with a probability of 50% at the end of each utterance. The DOAs of the two speakers are never the same, i. e., the minimum difference between the azimuth angles is $20°$ due to our RIR recording setup. Mixture SNR and source-array distance (same for both speakers) are indicated for each experiment. The recorded pub noise is added to the mixture according to the specified SNR, whereas the level of the ambient noise is fixed (negligible compared to the pub noise at $\mathrm{SNR} \lesssim 20\,\mathrm{dB}$).

First, we will focus on the problem of extracting estimates of both clean speech signals. For this purpose, we set the center direction of the target DOA range $\varphi_{\mathrm{c}}(t)$ to the estimate of the respective azimuth angle of arrival $\varphi_j(t)$. With a tolerated deviation of $\Delta\varphi = 10°$, the target source should still be extracted correctly in the presence of small DOA estimation errors, but the resolution remains sufficient to permit (at least theoretically) the separation of two sources that are only $20°$ apart. A dedicated evaluation to determine how φ_{c} and $\Delta\varphi$ can be chosen appropriately will be conducted in Section VI-D.

Unless otherwise specified, we do not assume knowledge of the source locations, but make use of our CNN/LSTM broadband DOA estimator from [34]. The network receives the microphone signal phase spectrograms as input, and returns frame-wise source presence probabilities for each DOA $\psi_d \in \Psi$. Oracle knowledge is only used to determine the *number* of speakers J and to assign the DOA estimates to the speakers (by selecting the permutation with the lowest sum of absolute DOA estimation errors).

From the introduced methods, we choose three for the evaluation. As discussed in Section III, the redundant generation of $D = 72$ outputs adversely affects the efficiency of the LBO approach, and [19] demonstrated that LBT outperforms PIT for a (multichannel) setup where spatial information is available. We therefore select the LBT of Section III-B (based on [19]) as a representative of the multi-output methods. Although knowledge of the source DOAs is technically not required, the allocation of the sources to the output channels can change abruptly when the source locations change. To obtain an upper bound for the performance of *this baseline only*, we use the true (oracle) DOAs to select the correct output channel in each frame. As a second baseline, we consider the LBI of (23), which can be seen as the application of [24] to the SSS problem (as discussed in Section IV-A). Finally, we consider the LDE of Section IV-B, which is the main contribution of this work.

Each of these three methods are tested based on both the LC and the HC version of CRUSE. An overview of the resulting numbers of trainable parameters and MAC operations per frame is displayed in Table I. To *separately* extract multiple sources with the SSS, the same input has to be processed with different target DOA ranges $\Psi^*(t)$, thereby raising the number of MACs per frame accordingly. Otherwise, the differences between LBT, LBI, and LDE are rather small.

The output signals are obtained by applying the masks as in (7). As quality metrics, we use the intrusive segmental target-to-noise ratio (TNR) and target-to-interferences ratio (TIR), as well as the (extended) short-time objective intelligibility (STOI) [39] and wideband perceptual evaluation of speech quality (PESQ) [40]. STOI values range from 0 to 1, whereas PESQ scores are between 1.02 and 4.56 on the MOS-LQO scale. To compute TIR and TNR, we exploit the fact that an estimated mask can be applied individually to all signal

components to analyze its effect. Residual target speech, interference (including reverberation), and noise (including ambient), as given by

$$S_{\mathrm{res}}(f,t) = \mathcal{M}(f,t)\, S(f,t) \tag{28a}$$

$$I_{\mathrm{res}}(f,t) = \mathcal{M}(f,t) \sum_{j \notin \mathcal{J}(t)} S'_j(f,t) + S''_j(f,t) \tag{28b}$$

$$V_{\mathrm{res}}(f,t) = \mathcal{M}(f,t)\, V(f,t), \tag{28c}$$

are first transformed back into the time domain, and then again divided into frames using the same frame length and shift. The microphone index is omitted to indicate that the reference channel is used (same convention as in (5)). We then define

$$\mathrm{TIR}\,[\mathrm{dB}] = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} 10 \log \frac{\sum_k s^2_{\mathrm{res}}(k,t)}{\sum_k i^2_{\mathrm{res}}(k,t)} \tag{29a}$$

$$\mathrm{TNR}\,[\mathrm{dB}] = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} 10 \log \frac{\sum_k s^2_{\mathrm{res}}(k,t)}{\sum_k v^2_{\mathrm{res}}(k,t)}, \tag{29b}$$

where $k \in \{0, \ldots, F-1\}$ indexes the samples within one frame, and lowercase letters denote the time domain counterparts of the respective STFT domain signals. Summation is performed over active frames $t \in \mathcal{T}$. A frame is considered active when its energy $\sum_k s^2_{\mathrm{res}}(k,t)$ exceeds 1% of the value at the 95th percentile of the frame energies of the entire signal.

B. Numerical Results

The results are shown in Fig. 8 as a function of the input SNR (first two columns) and the source-array distance (last two columns). To obtain meaningful numbers, each data point in the plot is the average score of 25 simulations conducted for one set of test parameters and both of the $J = 2$ sources.

1) Robustness: Generally, a good performance is also observed under challenging conditions, regardless of whether LBT, LBI, or LDE is used. The fairly consistent increase of the TIR and TNR scores over the input mixture (—|—) demonstrates that a comparable suppression of unwanted signal components is achieved across all considered input SNRs and distances. A similar trend is seen for STOI as well.

Note that the PESQ scores for the input mixtures are near the theoretical minimum of 1.02 in all cases due to the predominance of unwanted signal components, which include reverberation, interference, *and* noise. Around this lower end of the scale, even a large *subjective* improvement might only translate to small PESQ changes. Consequently, the achieved absolute scores are relatively low (maximum of 1.71). Also, as previously reported, e. g., in [41], directly applying masks as in (7) improves the PESQ metric less effectively as compared to using the same masks for mask-based beamforming. Nevertheless, we can here consider the relative differences as an indication of the achieved improvement. Note that the speech quality is also generally expected to improve when clean magnitude and phase are estimated jointly. However, we anticipate similar performance trends in this case, because this aspect is independent of the choice of LBT, LBI, or LDE.

2) Results for fixed vs. changing speaker locations: Comparing the first column (time-invariant speaker locations) with the second (time-variant), and the third column (time-invariant) with the fourth (time-variant), the differences in terms of the (segmental) TIR and TNR, as well as STOI, are relatively small. This confirms that considering fixed source and target DOAs during training is sufficient. The larger impact on PESQ is because this metric is more strongly affected by a *temporarily* reduced performance. When the locations of the two speakers are time-variant, the repeatedly changing difference between the corresponding azimuth angles is likely to be small (e. g., 20°, the most challenging case) *at least once* within each experiment. When the source locations are fixed, the performance is expected to be more consistent over the entire signal duration. Only for the LBT baseline, a somewhat larger effect of DOA changes is also observed in terms of the other metrics: for example, the STOI score of HC-LBT (—●—) drops from 0.59 to 0.53 for an input SNR of 10 dB and a distance of 2 m, whereas the STOI of HC-LDE (—◆—) only decreases from 0.64 to 0.61. For conciseness, we mainly focus on the setup with fixed locations in the following.

3) Comparison of the two baseline approaches: Leaving the architecture unchanged (LC or HC), we observe that LBI performs slightly better than LBT in terms of STOI and PESQ, even when LBT achieves a higher interference suppression (increased TIR). However, HC-LBT outperforms LC-LBI (--□--), suggesting that this minor improvement might not make up for the increased cost of extracting the sources individually (not required when using LBT with 2 outputs).

4) Proposed LDE vs. baselines: Already the LC-LDE results (--◇--) *at least match* the performance of the more computationally demanding HC setups of both baseline approaches, barring a slightly lower STOI at high input SNRs (for SNR = 30 dB and a 2 m distance: 0.66 for LBI and LBT, 0.64 for LDE). With HC-LDE (—◆—), we obtain results that clearly outperform all other considered approaches regarding STOI and PESQ, e. g., PESQ = 1.54 with HC-LDE for SNR = 10 dB and a 2 m distance, as compared to the best baseline where PESQ = 1.37. At the same time, TIR and TNR indicate that unwanted components are not suppressed significantly more strongly. Thus, it stands to reason that the improvement is primarily due to the masks capturing the target speech more accurately, thereby reducing its distortion. To verify this, we will now look at one example more closely.

C. Selected Example

We consider an example with a female *target* speaker and a male *interfering* speaker. For both, the impulse responses recorded for a source-array distance of 2 m are used, such that the reverberant target and interfering speech signals are approximately equally strong. As shown in the top right of Fig. 9, their DOAs change after each utterance. Thereby, we can also study the impact of the azimuth difference on the separation quality: the sources are far apart at first, then approach each other, and farther apart again in the end. The mixture SNR is set to 5 dB. The resulting spectrogram of the input signal is depicted in the top left. The remaining spectrograms

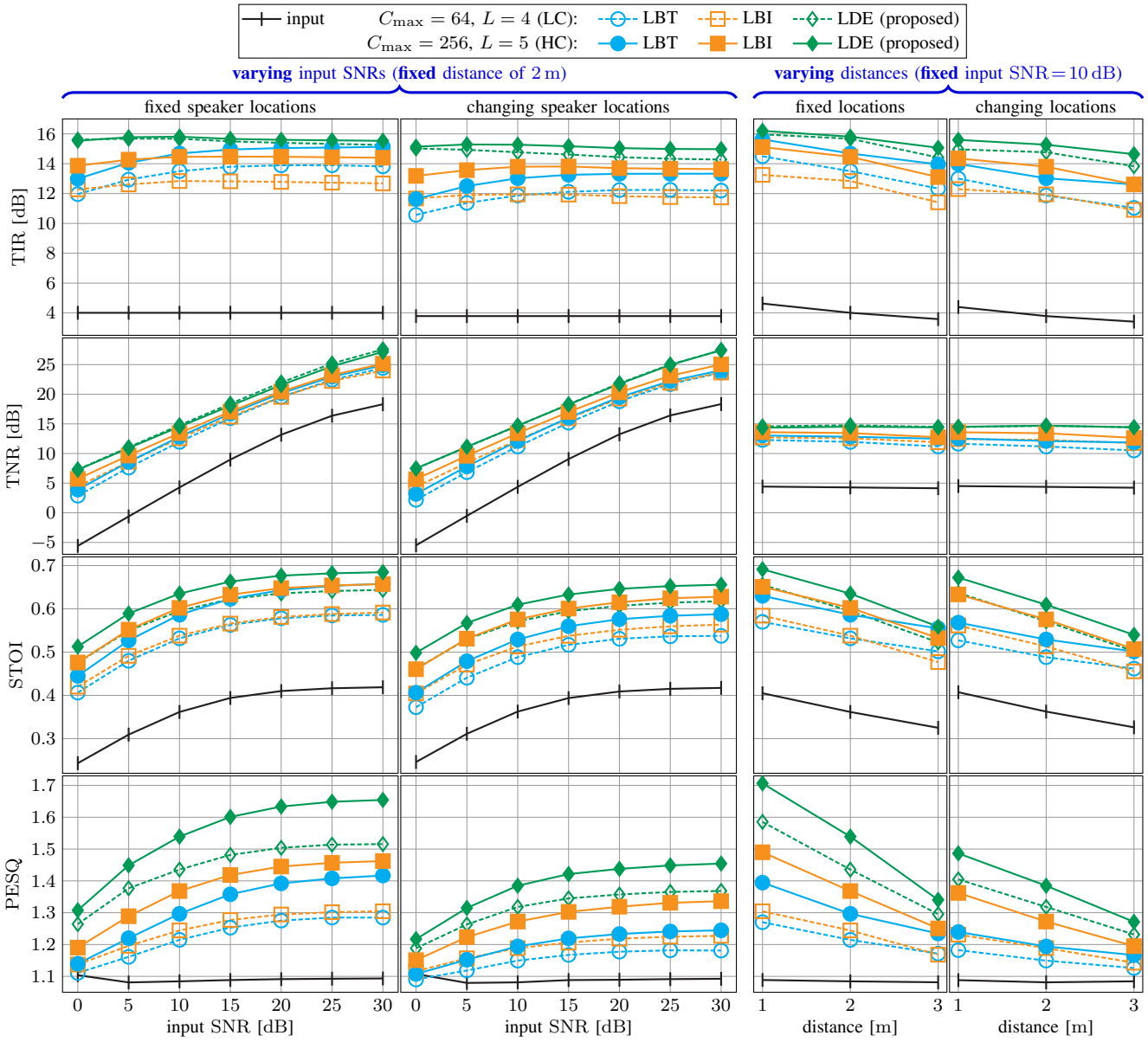

Fig. 8. Average scores based on both of $J = 2$ speakers in 25 simulations and as a function of the input SNR (left) and the source-array distance (right). We observe that the LC setup of the proposed LDE performs comparably as the HC setups of the considered baselines, and HC-LDE performs even better.

correspond to the clean speech signals (second row), as well as the output signals that the LC and HC architectures yield for the target speaker using the LBT, LBI, and LDE methods. Since the differences can be seen most clearly at lower frequencies, where the speech spectrogram is most distinct due to its harmonic structure, only frequencies up to $4.5\,\mathrm{kHz}$ are shown (the sampling rate is still $16\,\mathrm{kHz}$). To facilitate the interpretation of the results, we assume the availability of the oracle DOAs for this example, i. e., the center direction is set to $\varphi_{\mathrm{c}}(t) = \varphi_1(t)$ in all frames (see Section VI-D for a closer analysis of the influence that an offset between $\varphi_{\mathrm{c}}(t)$ and the source DOA has on the results). The setup is otherwise still as described in Section VI-A. All audio signals are available at https://aspire.ugent.be/demos/TASLP2023AB/.

In the LC-LBT spectrogram, the speech structure cannot be recognized very clearly, especially at low frequencies, which indicates that the estimated mask does not capture the *details* of the target signal with the desired accuracy. Unwanted components are not suppressed effectively near TF bins where the target is strong, e. g., between harmonic frequencies. This leads to a speech-like color of the residual noise that gives rise to a *perceived* target signal distortion in the resulting audio file. A clear improvement in this regard is achieved with LC-LBI and (even more so) LC-LDE. Whereas particularly the residual interference remains audible with LC-LBI, LC-LDE succeeds in producing a relatively clean output when the source DOAs are sufficiently far apart (60° or more). The target speech is captured very well at frequencies above $0.5\,\mathrm{kHz}$.

Upon comparing the LC and HC outputs, we note that the increased complexity generally enables a better performance across the entire spectrum. The main improvement, however, is observed at low frequencies, where HC captures the target

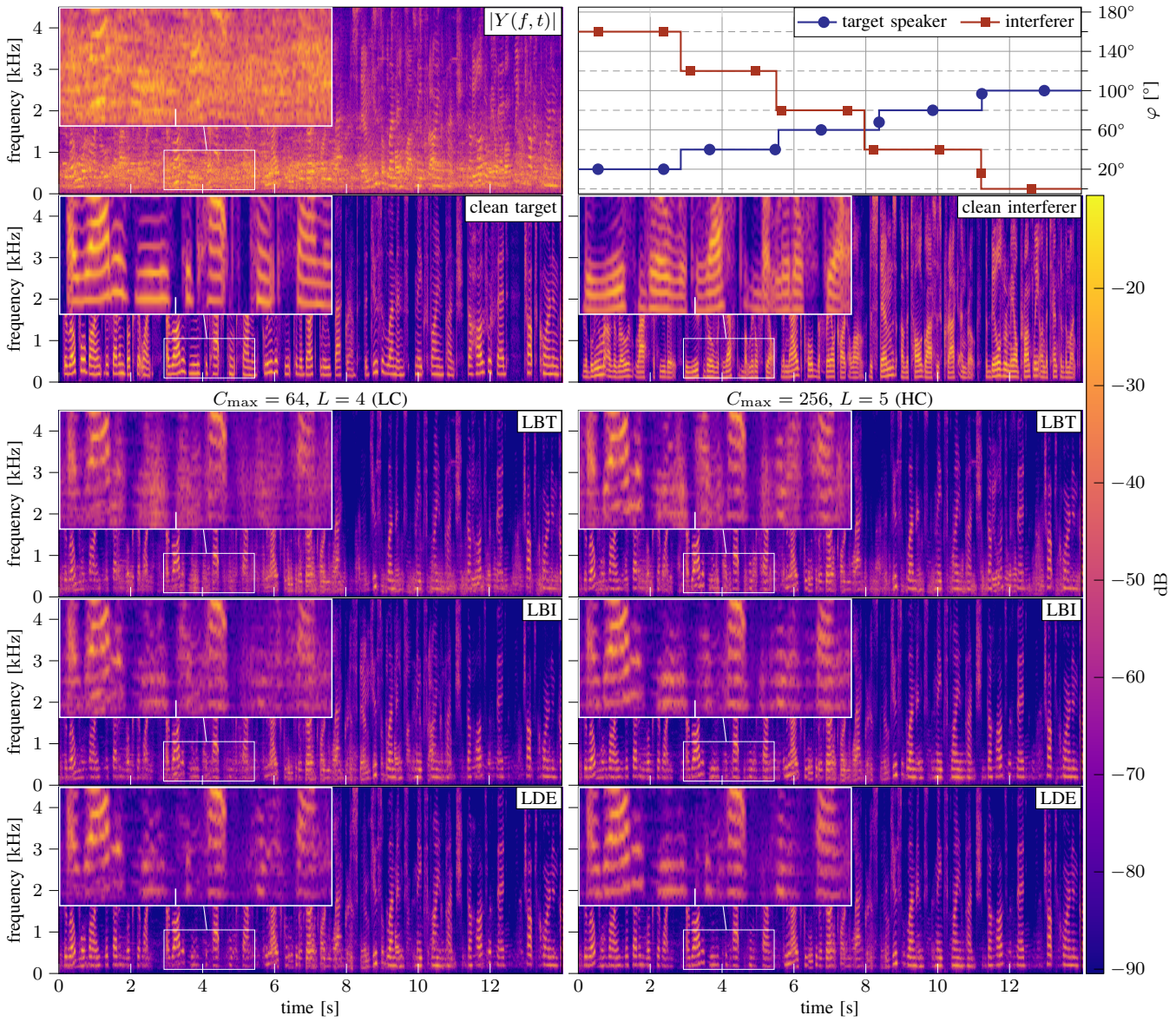

Fig. 9. Spectrograms for an example with one female (target) and one male (interfering) speaker, as well as additive noise with SNR $= 5\,\mathrm{dB}$. Source and target DOAs change after each utterance. Only frequencies up to $4.5\,\mathrm{kHz}$ are shown so that low frequency differences can be seen more clearly. We observe that the proposed LDE approach best captures the target signal, thereby suppressing unwanted components effectively while limiting speech distortion.

signal significantly better. As to the comparison of HC-LDE with the baseline approaches, we again note that the reduced distortion is the main advantage over LBT, and the more effective suppression of interference and noise is the main advantage over LBI. This is in line with the findings from Fig. 8, where LBI outperformed LBT in terms of PESQ despite the higher TIR scores of LBT, and LDE generally performed better than both baselines.

Considering the changing source locations, the interfering speaker is suppressed effectively with all methods whenever the azimuth angles differ by at least 60°. For the spacing of 40° (between $8.4\,\mathrm{s}$ and $11.2\,\mathrm{s}$), there remains a relatively significant residual interference *only* for the LC-LBI setup. Between $5.6\,\mathrm{s}$ and $8.0\,\mathrm{s}$, however, where the difference between the two azimuth angles is only 20°, the speakers are not separated successfully with either setup. It is interesting to note that the second LBT output channel then contains (almost) only noise, whereas the target and interferer estimates produced by LBI and LDE are relatively similar mixtures of both speakers.

These findings then raise the question what difference between the DOAs is required so that the sources can still be separated, which will be addressed in the next section. Further, we will discuss how the target DOA range $\Psi^*(t)$ can be selected appropriately, e. g., based on given DOA estimates.

D. Spatial Selectivity Analysis of the Proposed Approach

The goal of the location-based SSS is to suppress sources from nontarget directions $\psi_d \notin \Psi^*(t)$ while minimizing the distortion of sources from directions $\psi_d \in \Psi^*(t)$ (see Section IV). When (25) is used to define $\Psi^*(t)$ with a center direction φ_c and a tolerated deviation $\Delta\varphi$, this implies that

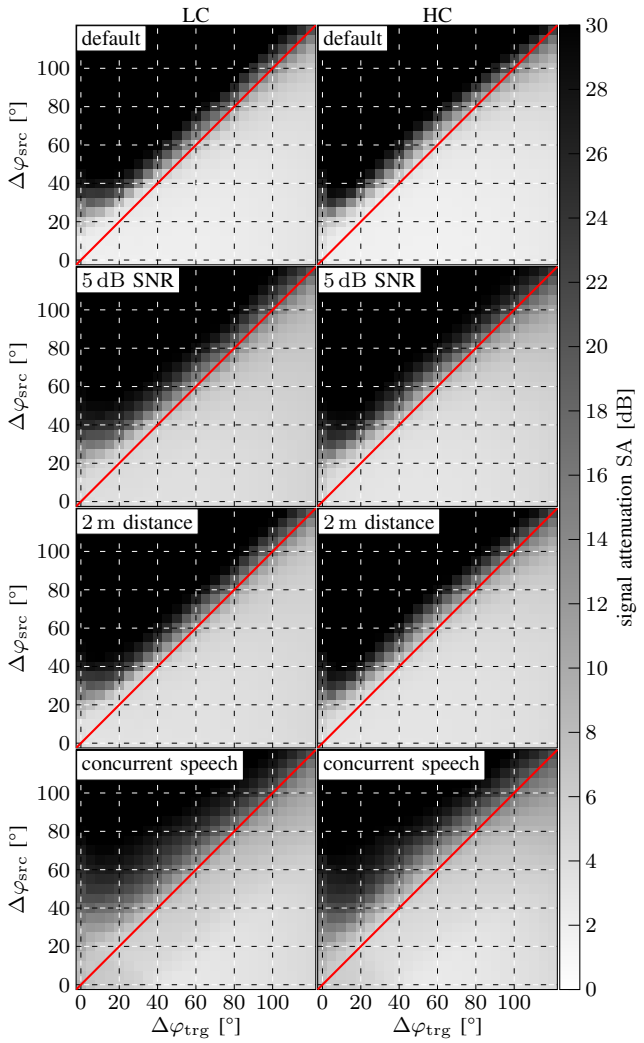

Fig. 10. Attenuation of a talker with DOA φ_1, when the *actual* offset to the center direction is $\Delta\varphi_{\mathrm{src}} = |\mathrm{angdif}(\varphi_1, \varphi_{\mathrm{c}})|$, and $\Delta\varphi_{\mathrm{trg}}$ is the *tolerated* offset. Deviations from the default setup (30 dB SNR, 1 m distance, no other talkers) are noted in each plot. Ideally, we expect a low attenuation (light color) below and a strong attenuation (dark color) above the main diagonal.

the direct-path component of a source located at a (time-invariant) azimuth angle φ_1 is preserved under the condition that $|\mathrm{angdif}(\varphi_1, \varphi_{\mathrm{c}})| \leq \Delta\varphi$. For clarity, we will denote the user-defined tolerated deviation by $\Delta\varphi_{\mathrm{trg}} = \Delta\varphi$, and the offset of the source DOA compared to the specified center direction by $\Delta\varphi_{\mathrm{src}} = |\mathrm{angdif}(\varphi_1, \varphi_{\mathrm{c}})|$ in this section. To verify that the use of LDE leads to this desired behavior, Fig. 10 shows how strongly a source is attenuated at different values of $\Delta\varphi_{\mathrm{src}}$ and $\Delta\varphi_{\mathrm{trg}}$. This may be compared to the (broadband) beampattern of a beamformer, where $2\Delta\varphi_{\mathrm{trg}}$ is the (adjustable) beamwidth and $\Delta\varphi_{\mathrm{src}}$ is the source DOA (relative to the steering direction). We consider the signal attenuation of this one source (with index $j = 1$)

$$\mathrm{SA}\,[\mathrm{dB}] = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} 10 \log \frac{\sum_k s^2(k,t)}{\sum_k s_{\mathrm{res}}^2(k,t)}, \tag{30}$$

for which we set $\mathcal{J}(t) = \{1\}$ (*only* to evaluate (30), i. e., we do *not* train a new model for this experiment) in the definition of $S(f,t)$. The corresponding residual signal $S_{\mathrm{res}}(f,t)$ is given by (28a), $s(k,t)$ and $s_{\mathrm{res}}(k,t)$ are the respective time domain signals. Note that (30) does not account for reverberation. For each pair of $(\Delta\varphi_{\mathrm{trg}}, \Delta\varphi_{\mathrm{src}}) \in \{0°, 5°, \ldots, 120°\}^2$, we performed 15 simulations and averaged the obtained SA values. One of the recorded RIRs is randomly selected each time, such that $\varphi_1 \in \{0°, 20°, \ldots, 180°\}$, and either of the two possible center directions $\varphi_{\mathrm{c}} \in \{\varphi_1 \pm \Delta\varphi_{\mathrm{src}}\}$ is chosen.

We first consider favorable conditions, where the input SNR is 30 dB, the source-array distance is 1 m, and there are no other talkers ($J = 1$). This serves as the *default setup* (first row) in Fig. 10. The results for the LC and HC architectures are displayed on the left and the right side, respectively. As expected, we observe that the signal is mostly suppressed when $\Delta\varphi_{\mathrm{src}} > \Delta\varphi_{\mathrm{trg}}$ (i. e., actual offset exceeds tolerated offset), and preserved relatively well when $\Delta\varphi_{\mathrm{src}} \leq \Delta\varphi_{\mathrm{trg}}$.

Around the edge of the target DOA range ($\Delta\varphi_{\mathrm{src}} \approx \Delta\varphi_{\mathrm{trg}}$), there is a transition region where the signal is partially suppressed. Interestingly, this transition only really starts outside the target DOAs (above the main diagonal), presumably because the training focuses on estimating the target signal, whereas interferers are only implicitly accounted for. The suppression really becomes effective ($\mathrm{SA} > 20\,\mathrm{dB}$) when the source is at least 20° outside of the target range ($\Delta\varphi_{\mathrm{src}} \geq \Delta\varphi_{\mathrm{trg}} + 20°$). One notable exception is that sources are never suppressed effectively when $\Delta\varphi_{\mathrm{src}} < 30°$. Thus, to ensure that all interferers with $\Delta\varphi_{\mathrm{src}} \geq 30°$ can be removed without unnecessarily distorting the target signal, a value of $\Delta\varphi_{\mathrm{trg}} = 30° - 20° = 10°$ should be selected based on the above two observations. Note that these findings may vary depending on, e. g., training setup and array geometry.

Next, we repeat this experiment with a reduced SNR (second row), a larger distance (third row), and an additional concurrent speaker (azimuth φ_2) where $|\mathrm{angdif}(\varphi_1, \varphi_2)| = 40°$ (fourth row). Compared to the original setup of the first row, raising the noise level or the distance does not drastically change the results, except for the growing difficulty of reliably separating the different signal components: the SA increases for target sources with $\Delta\varphi_{\mathrm{src}} \ll \Delta\varphi_{\mathrm{trg}}$ but decreases for non-target sources with $\Delta\varphi_{\mathrm{src}} \gg \Delta\varphi_{\mathrm{trg}}$, and the transition around $\Delta\varphi_{\mathrm{src}} \approx \Delta\varphi_{\mathrm{trg}}$ becomes slightly less sharp. This is expected in more challenging conditions because of the increasingly compromised approximate disjointness of the signal components, and due to the growing uncertainty regarding the source locations. A second (concurrent) speaker additionally causes the offset needed for an effective suppression to increase: even with the HC architecture, an attenuation of more than 20 dB is then only achieved at $\Delta\varphi_{\mathrm{src}} \geq 40°$. This could be a result of training the DNN to extract a *variable* (unknown) number of target sources instead of optimizing specifically for the task of extracting a single speaker.

Finally, when comparing the results for LC (left column) and HC (right), we see that differences in terms of the SA scores are relatively minor. This suggests that, although the choice of the network size has a considerable influence on the speech quality, the spatial selectivity is *primarily* determined by the design of LDE and training data, as well as the limited reliability of spatial information when an array with only 3 microphones is used. However, upon looking more closely

at the setup with concurrent speech (last row), it can be noted that the HC network still achieves a slightly increased attenuation of a speaker who is located just outside of a narrow target DOA range (small value of $\Delta\varphi_{\mathrm{trg}}$): for example, we obtain $\mathrm{SA} = 15\,\mathrm{dB}$ with HC, but only $12\,\mathrm{dB}$ with LC for $\Delta\varphi_{\mathrm{src}} = 30°$ and $\Delta\varphi_{\mathrm{trg}} = 10°$. This shows that the use of the more complex network also benefits the separation of speakers with similar DOAs.

E. Application to Moving Speaker Recordings

To better understand the implications of the observed spatial selectivity for the performance in practice, we now use HC-LDE to process recordings of moving talkers. A mixture of two talkers (both male) is considered. By recording them *separately*, we obtain a reference for the "clean" signals (where reverberation and background noise are still present, but not the interfering speaker). Relatively accurate ground truth information on the speaker DOAs was extracted from these clean signals using the approach of [42]. The two talkers moved continuously around a table in the meeting room described in Section VI-A while reading from the Harvard sentences [43].

The magnitude spectrograms of the two speakers are presented in the top left of Fig. 11 using two different colors (blue and red). The input to the separation network is the mixture signal (sum of these two recordings). As seen from the DOAs in the top right subplot, the speakers switch between clockwise and counterclockwise movement repeatedly, and their paths cross twice (once after $7.5\,\mathrm{s}$ and once after $21\,\mathrm{s}$).

In the left column of Fig. 11, the middle two rows each depict the spectrogram of *one* of the recorded speech signals. First, we consider the problem of extracting either of the two talkers. This is done by choosing $\varphi_{\mathrm{c}}(t) = \varphi_j(t)$ with $j \in \{1, 2\}$, and $\Delta\varphi = 10°$. We again assume oracle knowledge of the DOAs to simplify the interpretation of the results. Although it would be difficult to accurately localize the two talkers while the difference between their azimuth angles is small, a diminishing separation quality is anyway expected while both speakers are inside or near the defined target spatial region.

The output may be seen as the mixture of the masked first speaker and the masked second speaker (i. e., the signals we obtain by applying the same mask to the individual speech signals), where the two components are again represented by blue and red colors in the right column of the figure. This way, we can see how the second speaker is attenuated quite effectively when the first speaker is the target (blue color is prevalent) and vice versa, except during times where the two speakers have similar DOAs (i. e., around the time where they cross paths). This is also verified by the audio files available at https://aspire.ugent.be/demos/TASLP2023AB2/.

To demonstrate the usefulness of the spatially selective speaker separation beyond the decomposition of the mixture into the individual speech signals, we can also choose a fixed DOA range, whereby the target speaker changes over time. The yellow colored region in the DOA plot of Fig. 11 (top right) shows $\Psi^*(t)$ for $\varphi_{\mathrm{c}} = 180°$ and $\Delta\varphi = 35°$. In this case, the second speaker is briefly inside the target region at the beginning, followed by a longer period of time where the first speaker is the target. At the end, only the second speaker is inside the target region once more.

The corresponding outputs (oracle separation and estimate) are shown in the last row of the figure. Again, we observe that while the DOA of one of the speakers is within the defined target range, the corresponding speech signal is extracted correctly. However, the speaker is not immediately suppressed upon leaving the target region. This is in line with the findings of Section VI-D, which indicated that the suppression increases gradually as the distance from the boundary of the target DOA range increases.

VII. LibriCSS Continuous Input Evaluation

The results of Section VI-E demonstrate that the LDE-based system is effective when applied to real recordings. We now consider LibriCSS [26] to further verify our findings regarding the comparison of LDE with the LBI and LBT baselines. The LibriCSS dataset was created by recording utterances reproduced by loudspeakers at different locations in a meeting room. For each recording, an (approximate) ratio of overlap was defined, as compared to our experiments with fully overlapping speech in Section VI. The array consisted of 6 microphones uniformly spaced on a circle of radius $4.25\,\mathrm{cm}$ and an additional microphone at the center ($N = 7$). To evaluate the performance, word error rate (WER) scores are computed for the outputs generated by the ASR system provided with LibriCSS.

Due to the different array geometry, we retrained the localization and separation networks for this experiment. Otherwise, the training setup is still as described in Section V.

As [26] reports for the reference system of [44], significantly better ASR results may be achievable with mask-based minimum variance distortionless response (MVDR) beamforming. To estimate the signal statistics, ideally, interference and noise masks should then be acquired in addition to the one for the target. Alternatively, magnitude and phase could be enhanced jointly by CSM [27]. For simplicity, we only consider the signals obtained by multiplication with the masks (as in (7)) as input to the ASR system here, as this requires no changes to our SSS setup where only an estimate of the target IRM is returned. The resulting scores are used to compare LBT, LBI, and LDE, using either the LC or the HC network.

Again, φ_{c} is chosen to be the estimated DOA of the speaker we want to extract, and the width of the target region is fixed with $\Delta\varphi = 10°$. During speech inactivity (as determined based on the speech activity probabilities obtained from the DOA estimator), $\Psi^* = \emptyset$, whereby a silent output is generated. Despite the LBI and LDE results being dependent on the quality of the DOA estimator in general, the comparison remains fair as the same DOAs are used for both.

For LBT, we leave the order of the 2 output channels unchanged, but still use the output of the DOA estimator to *count* the number of currently active speakers. During inactivity of an output channel, the corresponding signal is replaced by silence.

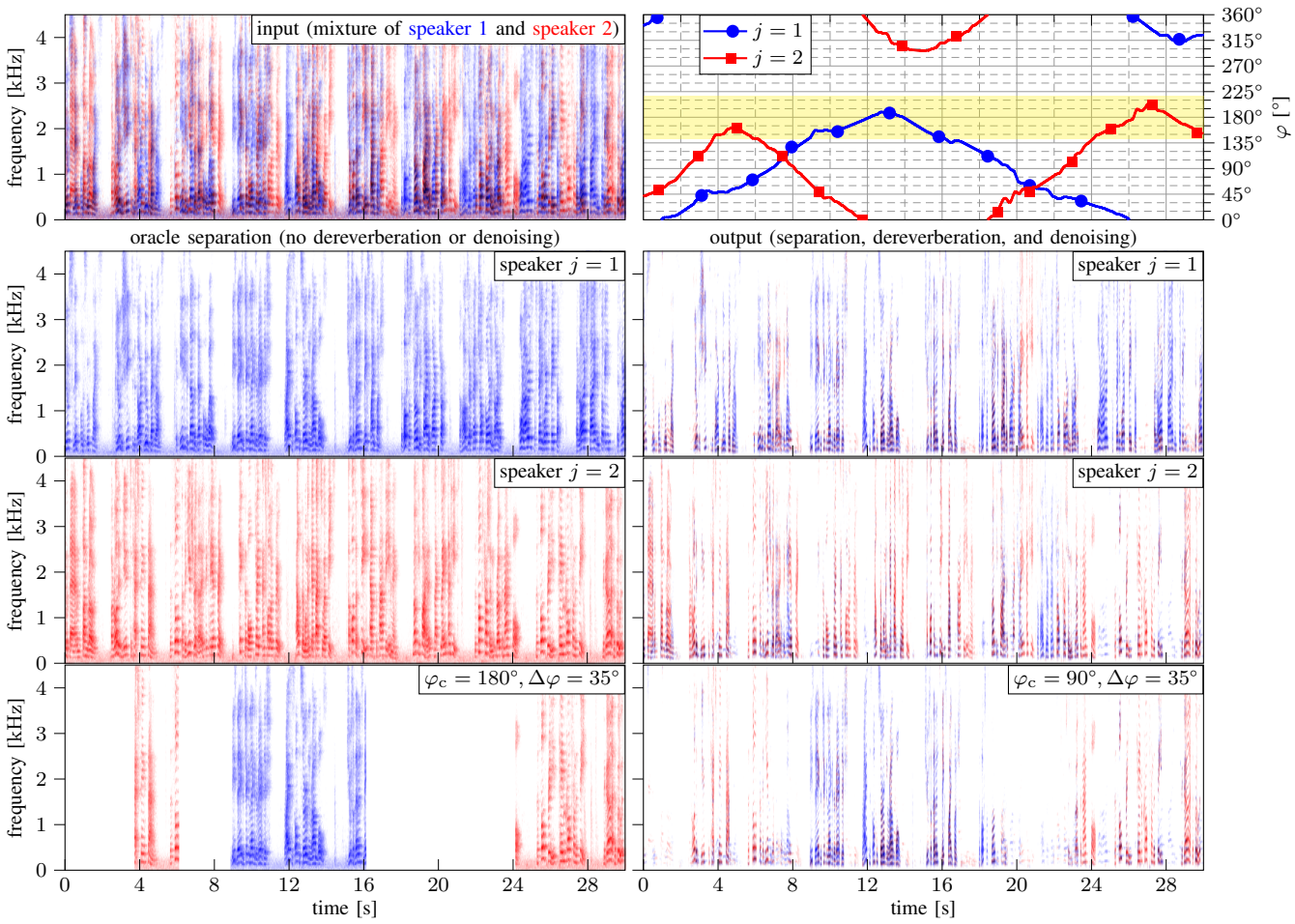

Fig. 11. HC-LDE is applied to a mixture of two separately recorded moving talkers (represented by blue and red colors). Whereas individual speakers are extracted in the two middle rows (target region based on the respective speaker DOA), in the bottom row we choose a fixed target region (yellow colored area in the DOA plot) which is independent of the speaker locations.

TABLE II
LIBRICSS CONTINUOUS INPUT EVALUATION. WHEREAS THE LIBRICSS REFERENCE SYSTEM USES MVDR BEAMFORMING, WE ONLY APPLIED MAGNITUDE MASKING AND KEPT THE NOISY PHASE FOR OUR RESULTS.

WER [%]		overlap ratio in %					
		0S	0L	10	20	30	40
no separation		15.4	11.5	21.7	27.0	34.3	40.5
reference [26], [44]		11.9	**9.7**	13.6	**15.0**	19.9	21.9
LC-	LBT	16.0	18.3	18.8	22.7	25.5	28.5
HC-	LBT	14.4	17.1	16.7	20.0	22.1	24.3
LC-	LBI	12.9	13.7	17.1	22.3	27.0	29.7
HC-	LBI	11.3	11.8	14.7	18.7	22.9	23.8
LC-	LDE	11.3	11.2	13.6	16.3	19.7	20.8
HC-	LDE	**10.8**	11.1	**12.8**	**15.0**	**17.9**	**18.9**

The results can be found in Table II. Reference scores for no separation and for the baseline of [44] were adopted from [26] (as LibriCSS "baseline", we here consider the 7-channel results of Table 3 in [26]). The speaker overlap is indicated for each column. Further, the "S" and "L" suffixes in 0S and 0L, respectively, indicate short and long silences between utterances.

Overall, we find that the LDE results are quite convincing for conditions with overlapping speech, considering that we only performed magnitude masking and did not use any context from future frames. Aside from this, we observe similar trends as in Section VI-B: LDE outperforms LBT and LBI for both the HC and LC networks, with the difference increasing with the overlap ratio (for 40% overlap: 18.9% WER with HC-LDE, 23.8% with HC-LBI). Owing in part to the smaller difference between the performance of the HC network and the LC network in the case of LDE as compared to LBT and LBI, we obtain lower WERs with LC-LDE than with the more computationally costly HC-LBI and HC-LBT.

VIII. CONCLUSIONS

We studied the SSS problem, where the aim is to extract one or more speakers from a mixture with concurrent speakers and background noise. Unambiguously defining the expected output of a DNN is not trivial in this case, as additional information is needed to distinguish target and interfering speech. In a multichannel setup, spatial information can be used for this purpose. The recently proposed LBT separates all speakers simultaneously, but arranges the output channels according to the source locations for an improvement over the arbitrary permutation of PIT. When the expected number of concurrent sources can vary, however, it may be undesirable to associate

each output channel with one source. For a more focused approach, the target definition can instead already be narrowed down at the input. In prior work, this has been achieved by the use of location-based input features (LBI), which we extended, here, to the more general SSS problem. We then hypothesized that simply representing location information and microphone signals by separate input features which are jointly processed might not allow for a powerful location conditioning, as these are different types of information. This led us to propose a novel location dependent feature extraction (LDE), which forms the main contribution of this work. The underlying idea is to process the microphone signals in a location dependent manner, where the DNN parameters are independently optimized for each DOA. A range of target directions can be selected and readjusted dynamically, e. g., by the user, or based on source DOAs estimated beforehand. We experimentally demonstrated that LDE achieves a better trade-off between performance and computational complexity than the LBI and LBT baselines. Whereas all methods suppress unwanted components effectively, the LDE output captures the details of the target spectrogram most accurately, which results in the best speech quality. We also verified that the DNN learns the relation between the target DOAs and the corresponding expected output quite well, and used the findings to determine how these DOAs can be selected appropriately.

A notable limitation of the current approach is that sources with a DOA just outside of the specified target range are not suppressed very well. With the considered experimental setup, a difference of at least 40° between the azimuth angles was needed to ensure that target and interfering speakers are separated effectively. In future work, we aim to further refine how the LDE is performed to improve the spatial selectivity, e. g., by additionally extracting information about *unwanted* directions, which may be compared to the steering of nulls in classical beamforming. It may also be of interest to consider the source-array distance as an additional criterion, as [20] proposed for LBT.

Acknowledgments

This work is partially supported by the Research Foundation - Flanders (FWO) under grant numbers 11G0721N and G081420N.

References

[1] N. Kanda, C. Boeddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, and R. Haeb-Umbach, "Guided source separation meets a strong ASR backend: Hitachi/paderborn university joint investigation for dinner party ASR," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1248–1252.

[2] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2020, pp. 1–7.

[3] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 380–390, 2020.

[4] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 656–660.

[5] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6394–6398.

[6] J. Wu, Z. Chen, J. Li, T. Yoshioka, Z. Tan, E. Lin, Y. Luo, and L. Xie, "An End-to-End Architecture of Online Multi-Channel Speech Separation," in *Proc. 21st Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 81–85.

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

[8] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 1, 2002, pp. I–529–I–532.

[9] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 1562–1566.

[10] J. Du, Y. Tu, L.-R. Dai, and C.-H. Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 8, pp. 1424–1437, 2016.

[11] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 31–35.

[12] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 246–250.

[13] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 241–245.

[14] 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. Audio, Speech, Language Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[15] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 1–5.

[16] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5739–5743.

[17] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, and S. Gannot, "Multi-microphone speaker separation based on deep DOA estimation," in *Proc. 27th Eur. Signal Process. Conf.*, 2019, pp. 1–5.

[18] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Neural networks using full-band and subband spatial features for mask based source separation," in *Proc. 29th Eur. Signal Process. Conf.*, 2021, pp. 346–350.

[19] H. Taherian, K. Tan, and D. Wang, "Location-based training for multi-channel talker-independent speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 696–700.

[20] ——, "Multi-channel talker-independent speaker separation through location-based training," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 2791–2800, 2022.

[21] P. Pertilä and J. Nikunen, "Distant speech separation using predicted time–frequency masks from spatial features," *Speech Communication*, vol. 68, pp. 97–106, 2015.

[22] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 558–565.

[23] M. Yu, X. Ji, B. Wu, D. Su, and D. Yu, "End-to-End Multi-Look Keyword Spotting," in *Proc. 21st Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 66–70.

[24] S. Kindt, A. Bohlender, and N. Madhu, "Improved separation of closely-spaced speakers by exploiting auxiliary direction of arrival information within a U-Net architecture," in *Proc. 18th IEEE Int. Conf. Adv. Video Signal-Based Surveillance*, 2022, pp. 1–8.

[25] K. Tesch and T. Gerkmann, "Spatially selective deep non-linear filters for speaker extraction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[26] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7284–7288.

[27] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1778–1787, 2020.

[28] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. 19th Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 3229–3233.

[29] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. 24th Eur. Signal Process. Conf.*, 2016, pp. 1153–1157.

[30] Y. Yu, W. Wang, and P. Han, "Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks," *EURASIP J. Audio, Speech, Music Process.*, no. 1, pp. 1–18, 2016.

[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, D. N. L., and Z. V., "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1," *Linguistic Data Consortium*, 1993.

[32] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 1509–1512.

[33] E. A. P. Habets, "RIR generator," https://github.com/ehabets/RIR-Generator, accessed: August 04, 2022.

[34] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Exploiting temporal context in CNN based multisource DOA estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1594–1608, 2021.

[35] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Amer.*, vol. 122, no. 6, pp. 3464–3470, 2007.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representat.*, 2019, pp. 1–19.

[37] European Telecommunications Standards Institute, "Speech processing, transmission and quality aspects (STQ); speech quality performance in the presence of background noise; part 1: Background noise simulation technique and background noise database," ETSI EG 202 396-1, 2005.

[38] P. Kabal, "TSP speech database," McGill University, Montreal, Quebec, Canada, Tech. Rep., 2002.

[39] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[40] International Telecommunication Union, "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," ITU-T Recommendation P.862.2, 2007.

[41] L. Drude, D. Hasenklever, and R. Haeb-Umbach, "Unsupervised training of a deep clustering model for multichannel blind source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 695–699.

[42] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Least-squares DOA estimation with an informed phase unwrapping and full bandwidth robustness," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 4841–4845.

[43] E. H. Rothauser, W. D. Chapman *et al.*, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, no. 3, pp. 225–246, 1969.

[44] T. Yoshioka, H. Erdogan, Z. Chen, X. Xiao, and F. Alleva, "Recognizing Overlapped Speech in Meetings: A Multichannel Separation Approach Using Neural Networks," in *Proc. 19th Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 3038–3042.