WEAKLY DOA GUIDED SPEAKER SEPARATION WITH RANDOM LOOK DIRECTIONS AND ITERATIVELY REFINED TARGET AND INTERFERENCE PRIORS

Alexander Bohlender[1], *Ann Spriet*[2], *Wouter Tirry*[2], *Nilesh Madhu*[1]

[1] IDLab, Department of Electronics and Information Systems, Ghent University - imec, Belgium
[2] Goodix Technology (Belgium) B.V., Leuven, Belgium

ABSTRACT

Given a mixture of multiple speech signals, a neural network can extract the talkers individually and sequentially, which may improve the output quality compared to a simultaneous separation of all speakers. To still suppress interfering speech effectively, the residual mixture of the remaining unseparated speakers can be included in the input of the next step. We build upon this approach with a twofold contribution. First, we propose to refine the already extracted speech signals in further optional iterations. This is accomplished by exploiting that the outputs of previous steps provide prior information on interference and target. Experiments indicate a gradual improvement until convergence after about 2 iterations per speaker. Secondly, look directions are defined to control in which order speakers are extracted, thereby resolving the related permutation ambiguity. Whereas supplying the true speaker locations delivers the best results, even a weak guidance with random directions reduces interference leakage significantly.

Index Terms— step-wise speaker separation, weakly guided, look direction, microphone array, neural network

1. INTRODUCTION

Recorded speech signals often contain concurrent talkers, reverberation, and noise. A deep neural network (DNN) can extract the underlying clean signals by assigning every speaker to one output channel, but their order is arbitrary. For this reason, a permutation invariant training (PIT) [1, 2] is often used, where the DNN must learn to resolve the ambiguity itself. Given a mixture captured by a compact microphone array, spatial information presents an alternative. Sorting the speakers by direction of arrival (DOA) or distance permits a location-based training [3]. Uniquely defining the desired output in this manner enables an improvement over PIT.

On the other hand, experiments of [4, 5] indicate that focusing on speakers individually may improve the quality of the results. This can be accomplished by specifying a target DOA with the angle feature of [6] or a one-hot vector [7, 8]. We also proposed a location dependent feature extraction [5] for the task of isolating speakers within a defined spatial region. When the speaker locations are unknown, instead, a first subnetwork can provide separate information for all individual speakers, based on which a second subnetwork estimates the speech signals one by one. Such cascade models were considered in [2, 4, 9]. The recurrent selective hearing network of [10] extracts one speaker at a time by using the outputs of earlier steps to determine which speakers remain. For this purpose, the DNN receives an auxiliary input, e. g., the residual signal obtained either by subtracting the speech signals from the input mixture [9] or directly estimated [11].

This form of information from previous steps can potentially improve the separation. However, it remains arbitrary in which order speakers are extracted. To resolve this permutation ambiguity, we propose here to choose a *look direction* in every step. The target speaker is then the one whose DOA is closest to this look direction.

A step-wise separation is performed by estimating one speech signal with each forward pass of the DNN (referred to as one *iteration*). In this work, the iteration specific supplementary input is composed of up to three types of information. The first is an *interference embedding* obtained while other speakers were extracted in earlier iterations. Like the residual used in [9–11], its purpose is to describe the already separated speakers. The second is a *target embedding* obtained while the same speaker was extracted in an earlier iteration. This can help to improve the output quality in further (optional) iterations, as compared to prior work, where iterations were limited to one per speaker. Thirdly, a one-hot vector defines the look direction.

Experiments show a gradual improvement in the first 2 iterations per speaker, indicating effective use of target and interference information. Specifying look directions, even when they are chosen randomly and without knowledge of the speaker locations, reduces interference leakage. This way, we also come fairly close to the results obtained upon providing the true DOAs, although the margin increases when the speakers are located in close proximity to each other.

This paper is organized as follows: After Sec. 2 introduces the speaker separation problem, Sec. 3 explains the iterative framework established in prior work. The proposed approach is presented in Sec. 4 and then evaluated in Sec. 5. Sec. 6 concludes the paper.

2. PROBLEM FORMULATION

We consider the problem of estimating the clean speech signals from a mixture of J concurrent talkers captured by a compact array of N microphones. In the short-time Fourier transform (STFT) domain, the speech signals are represented by vectors $\mathbf{S}_j^{\mathrm{dr}}(t,f) \in \mathbb{C}^N$ with $\mathbf{S}_j^{\mathrm{dr}}(t,f) = \mathbf{S}_j^{\mathrm{d}}(t,f) + \mathbf{S}_j^{\mathrm{r}}(t,f)$, where $\mathbf{S}_j^{\mathrm{d}}(t,f)$ and $\mathbf{S}_j^{\mathrm{r}}(t,f)$ are, respectively, direct-path and reverberation, j indexes the speakers, t the frames, and $f \in \{0,\dots,F-1\}$ the discrete frequencies. With the additive noise $\mathbf{V}(t,f)$, which may include localized nonspeech sources, spatially diffuse noise, and sensor noise, the signal model is

$$\mathbf{Y}(t,f) = \sum_{1\le j\le J} \mathbf{S}_j^{\mathrm{dr}}(t,f) \; + \mathbf{V}(t,f). \tag{1}$$

A single-channel input signal $Y(t,f)$, which is needed to compute the output signal in Sec. 3, is defined by averaging the squared magnitudes of $\mathbf{Y}(t,f)$ and taking the phase of the first microphone:

$$Y(t,f) = \tfrac{1}{\sqrt{N}} \left\|\mathbf{Y}(t,f)\right\|_{\ell_2} \mathrm{e}^{\jmath \angle Y_1(t,f)}. \tag{2}$$

As target signals, we use the rescaled direct-path components

$$S_j(t,f) = S_j^{\mathrm{d}}(t,f) \cdot E\left\{ \frac{\left\|\mathbf{S}_j^{\mathrm{dr}}(t,f)\right\|_{\ell_2}}{\left\|\mathbf{S}_j^{\mathrm{d}}(t,f)\right\|_{\ell_2}} \right\}, \tag{3}$$

where $S_j^{\mathrm{d}}(t,f)$ is derived from $\mathbf{S}_j^{\mathrm{d}}(t,f)$ analogously to (2). The scaling compensates for the attenuation of (distant) sources. Expectation $E\{\cdot\}$ is approximated by averaging over all t and f.

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

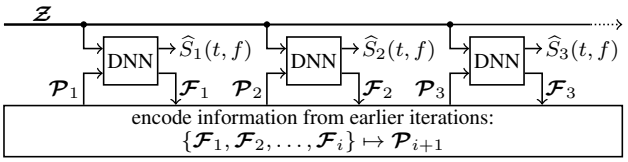


Fig. 1: Step-wise speaker separation (iterations from left to right). The DNN is the same in each iteration, only the input $\mathcal{P}_i$ (e. g., a residual mask) changes based on information $\mathcal{F}_i$ gathered previously.

3. DNN-BASED STEP-WISE SPEAKER SEPARATION

The recurrent selective hearing network proposed in [10] extracts speakers sequentially. A residual mask that indicates for each time-frequency bin which part of the mixture signal has not yet been allocated to a speaker is provided as auxiliary input. Later works like [9, 11, 12] instead use the residual signal to capture the information from previous iterations. Whereas this signal is obtained in [9, 12] by subtracting the already extracted speakers from the input mixture, [11] uses a separate output channel to estimate it directly.

In the following, $\mathcal{P}_i$ denotes prior information derived from features $\mathcal{F}_{i'}$ of previous iterations $i' < i$ as illustrated in Fig. 1. For example, $\mathcal{P}_i$ could be the residual signal when $\mathcal{F}_i$ is the estimated speech of one speaker. The input $\mathcal{Z}$ based on the microphone signals $\mathbf{Y}(t,f)$ remains the same in each iteration.

Based on our findings of [13], we opt for a hybrid masking and phase mapping in this work. The DNN returns three outputs, $M_i(t,f)$, $R_i(t,f)$, and $I_i(t,f)$, to compute the output signal

$$\widehat{S}_i(t,f) = 10^{M_i(t,f)} \, |Y(t,f)| \, \mathrm{e}^{\jmath \angle [R_i(t,f) + \jmath I_i(t,f)]}. \quad (4)$$

The ith iteration output (4) is an estimate of *either* of the J clean speech signals. This ambiguity was addressed in [10] with a variant of PIT [1, 2] allowing speakers to be extracted in an arbitrary order. However, studies like [3, 5, 8] show that using spatial information to *uniquely* define the desired output enables an improvement over PIT.

4. PROPOSED APPROACH

We propose a step-wise separation as in Fig. 1, where $\mathcal{P}_i$ combines *up to* three types of information: a target embedding $\mathcal{F}_i^{\mathrm{tar}}$, an interference embedding $\mathcal{F}_i^{\mathrm{int}}$, and a look direction $\mathcal{G}_i$. Sec. 4.1 introduces the DNN and discusses the choice of the latent features used as target and interference embeddings. The purpose of defining a look direction in Sec. 4.2 is to eliminate the need for PIT. The three types of prior information are jointly encoded by the network presented in Sec. 4.3.

4.1. Neural network architecture

Fig. 2 depicts the DNN applied once in each iteration. Barring the integration of $\mathcal{P}_i$ and the extraction of $\mathcal{F}_i$, it is identical to the convolutional recurrent U-Net for speech enhancement (CRUSE) of [14].

CRUSE is an encoder-decoder model, where L convolution layers form the encoder and L transposed convolution layers form the decoder. Convolutions along the time axis are applied in a causal manner with kernel size 2. Along frequency, the kernel size is 3 and the stride is 2, whereby the number of subbands is gradually reduced. The number of channels increases throughout the encoder with $C_\ell = \min\{2C_{\ell-1}, C_{\max}\}$ for $\ell \in \{2, \ldots, L\}$. Additive skip connections with learnable scaling and bias (in Fig. 2: represented by convolutions with $(1,1)$-kernel that are applied channel-wise) connect encoder and decoder. At the network bottleneck, the feature tensors are divided into 4 groups to reduce complexity when all subbands are jointly processed by parallel gated recurrent units (GRUs).

Here, the hyperparameters that control the trade-off between performance and complexity are $L = 5$, $C_1 = 64$, and $C_{\max} = 256$.

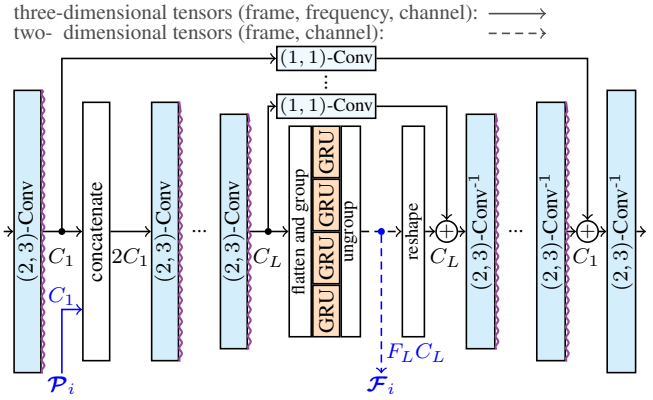


Fig. 2: U-Net based on [14]. Wavy lines (〰) indicate leaky ReLU. Prior information $\mathcal{P}_i$ is integrated after the first encoder layer. Latent features $\mathcal{F}_i$ carried over to later iterations come from the GRU output.

Along the channel axis, the entries of the input tensor $\mathcal{Z}$ are given by

$$\{\mathcal{Z}\}_{t,f} = \left[(\Re\{\mathbf{Y}'(t,f)\})^T, (\Im\{\mathbf{Y}'(t,f)\})^T \right]^T \in \mathbb{R}^{2N}, \quad (5)$$

where a causal moving average of all bins within the last 0.3 s is used for the normalization in $\mathbf{Y}'(t,f) = \mathbf{Y}(t,f) / E\{|Y(t,f)|\}$.

We choose the output of the GRU layer as $\mathcal{F}_i$. As the decoder is then no part of the loop, during inference it is only needed to obtain the output signals after the last iteration, which reduces the computational overhead of refining already extracted speech signals. By backpropagating the accumulated losses of several iterations, the use of latent features also makes it possible to take into account the effect that $\mathcal{F}_i$ has on later iterations in the optimization. However, e. g., the output of the last decoder layer, the resulting speech signal, or the residual signal remain equally suitable alternatives for $\mathcal{F}_i$. Comparing these different options is beyond the scope of this paper.

Prior information $\mathcal{P}_i$ is integrated by concatenation after the first encoder layer. We set the number of channels of $\mathcal{P}_i$ to coincide with $C_1 = 64$, which doubles the input size of the second encoder layer.

Dimensions: The embedding $\mathcal{F}_i$ has two dimensions, of which the first is the time axis. The second dimension has size $F_L C_L$, where F_ℓ denotes the number of subbands after $\ell \in \{1, \ldots, L\}$ encoder layers. The three-dimensional tensor $\mathcal{P}_i$ has the same shape as the output of the first encoder layer.

4.2. Weak DOA guidance

To let the DNN focus on a desired speaker, an additional input can be provided to indicate the corresponding (true or estimated) DOA. Such information is effectively encoded as a one-hot vector [7, 8] of which each entry represents one direction $\varphi \in \Phi$ of a discrete grid $\Phi = \{\varphi_1, \ldots, \varphi_D\}$. We refer to this as *strong DOA guidance*.

Here, we propose a *weak DOA guidance* with look directions $\varphi_i^* \in \Phi$ that can be chosen independently of the actual DOAs. The desired speaker is then the one whose DOA is closest to the current look direction (excluding already separated speakers). An example illustrates this in Fig. 3. Even a *random* look direction is sufficient to resolve the speaker permutation ambiguity in most cases (unless the DOAs of two speakers are equally far from the look direction).

Dimensions: The DOA guidance is provided by a two-dimensional tensor $\mathcal{G}_i$, where $\{\mathcal{G}_i\}_t$ is a one-hot vector of size D for each frame t.

4.3. Encoding of prior information

When $\mathcal{F}_{i'}$ is interpreted as a *speaker embedding*, it can take the role of an *interference embedding* $\mathcal{F}_i^{\mathrm{int}}$ when extracting a different speaker in a later iteration $i > i'$ or the role of a *target embedding* $\mathcal{F}_i^{\mathrm{tar}}$ when the aim is to refine the estimate of the same speaker. To

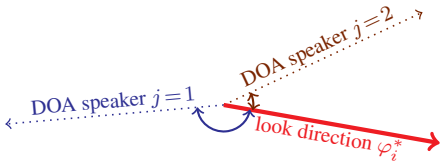


Fig. 3: For any look direction φ_i^*, the target speaker is the one whose DOA is closest (here: $j = 2$). No knowledge of the DOAs is needed.

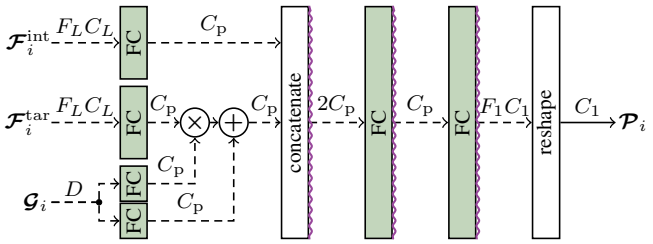


Fig. 4: The target speaker embedding $\boldsymbol{\mathcal{F}}_i^{\text{tar}}$, the interference embedding $\boldsymbol{\mathcal{F}}_i^{\text{int}}$, and the look direction $\boldsymbol{\mathcal{G}}_i$ are combined to form $\boldsymbol{\mathcal{P}}_i$.

extract speaker $j(i) = (i-1 \mod J) + 1$ in the ith iteration (i. e., to alternate between speakers cyclically), a suitable choice is thus

$$\boldsymbol{\mathcal{G}}_i = \boldsymbol{\mathcal{G}}_{i-J}, \quad i > J \tag{6a}$$

$$\boldsymbol{\mathcal{F}}_i^{\text{tar}} = \boldsymbol{\mathcal{F}}_{i-J} \tag{6b}$$

$$\boldsymbol{\mathcal{F}}_i^{\text{int}} = \max_{i-J<i'<i} \boldsymbol{\mathcal{F}}_{i'}, \quad J > 1, \tag{6c}$$

where $\boldsymbol{\mathcal{F}}_i = \mathbf{0}$ for $i \leq 0$, $\boldsymbol{\mathcal{F}}_i^{\text{int}} = \mathbf{0}$ when $J = 1$, and $\boldsymbol{\mathcal{G}}_i$ is defined for $i \leq J$ by selecting one unique look direction for every speaker. A stop flag [10] to indicate when all speakers have been separated could be straightforwardly introduced to cope with an unknown number of speakers J. For simplicity, however, we here assume J known.

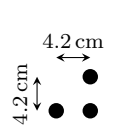
The three types of prior information must be fused to form $\boldsymbol{\mathcal{P}}_i$. Here, this is accomplished with the simple network of several fully connected (FC) layers depicted in Fig. 4, of which the hyperparameters were tuned empirically during development. First, target and interference embeddings are separately condensed to $C_\text{p} = 256$ features. To reflect that $\boldsymbol{\mathcal{F}}_i^{\text{tar}}$ and $\boldsymbol{\mathcal{G}}_i$ both characterize the *target*, they are combined by feature-wise linear modulation [15], i. e., by both multiplication and addition, which requires deriving two sets of C_p features from $\boldsymbol{\mathcal{G}}_i$. Target and interference features are then concatenated, so that the two following layers (C_p and $F_1 C_1$ output features, respectively) can process all information jointly. Finally, reshaping yields a three-dimensional tensor of the expected size.

5. EVALUATION

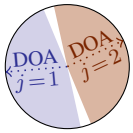
In Sec. 4, we proposed a weakly DOA guided speaker extraction, and discussed how the initial results could be refined in further iterations once all speakers are separated. In the following experiments, the focus lies on evaluating these key contributions of the paper.

Fig. 5a shows the considered array ($N = 3$ microphones). For each speaker, a clean signal is convolved with a room impulse response (RIR) in the time domain. Reverberant speech and noise are then added to obtain mixture signals. Different datasets are used for training and to test the trained models as detailed in Sec. 5.1 and 5.2.

The sampling rate is 16 kHz. For the STFT, we choose a frame length of 512 samples (32 ms), a shift of 160 samples (10 ms), and a square-root Hann analysis window. The DFT size of 512 implies that $F = 257$, $F_1 = 128$, $F_2 = 63$, $F_3 = 31$, $F_4 = 15$, and $F_5 = 7$. Synthesis is performed such that perfect reconstruction is possible. To define a grid for the look directions, considering only the azimuth angle, we set $\Phi = \{0°, 5°, \ldots, 355°\}$ ($D = 72$). For this setup, the U-Net has 7.0 M parameters and requires 44 M multiply-accumulate



(a) array geometry



(b) possible look directions (training)

Fig. 5: (a) Placement of the $N = 3$ microphones. (b) From both of the $J = 2$ shaded areas, one speaker specific look direction is selected.

(MAC) operations per frame (26 M without decoder). The network of Fig. 4 has 3.2 M parameters (3.2 M MACs).

5.1. Training setup

In the online data generation, first, one of 10 shoebox rooms and 7 array positions is selected, for which RIRs were simulated with [16] in advance. Within this room, the time-invariant speaker locations are drawn from the $D = 72$ azimuths of the grid Φ and 4 source-array distances. The number of speakers is $J \in \{1, 2\}$ (equal probability). Clean speech is taken from the TIMIT [17] and PTDB-TUG [18] corpora without rescaling the signals. A spherically isotropic noise field is simulated according to [19] to generate spectrally white but spatially diffuse noise. The SNR, given by the ratio of all reverberant speech to noise, is randomized within the range of 0 dB to 30 dB. Note that localized (nonspeech) noise sources could be additionally simulated to ensure robustness, but this is not considered here.

One look direction is assigned to each speaker. As illustrated in Fig. 5b, we randomly pick any direction that is "significantly" (threshold 10°) closer to the DOA of one particular speaker than to all others. This is done to avoid ambiguously defined training targets.

Batches consist of ten 2 s long mixtures. Based on the loss function of [14], omitting the t and f indices, we define

$$\begin{aligned} L_i = {} & \tfrac{1}{2}\,\text{MSE}\left[|S_{j(i)}|^c, |P\{\widehat{S}_i\}|^c\right] + \\ & \tfrac{1}{2}\,\text{MSE}\left[|S_{j(i)}|^c \mathrm{e}^{\jmath \angle S_{j(i)}}, |P\{\widehat{S}_i\}|^c \mathrm{e}^{\jmath \angle P\{\widehat{S}_i\}}\right], \end{aligned} \tag{7}$$

where $\text{MSE}[\cdot]$ is the mean square error, $P\{\cdot\} = \text{STFT}\{\text{ISTFT}\{\cdot\}\}$, and $c = 0.3$. The losses L_i of all iterations up to $i = 3J$ are accumulated before backpropagation. The learning rate of the AdamW optimizer is $8 \cdot 10^{-5}$ and the weight decay is 0.1.

5.2. Experimental setup

Each source signal is a root mean square normalized concatenation of 5 utterances from the TSP database [20] spoken by the same person. Sources are placed at azimuths $\varphi \in \{0°, 20°, \ldots, 180°\}$ in 1 m or 2 m distance from the array, for which we measured RIRs in two meeting rooms with similar reverberation times (slightly above 0.6 s). Although the proposed approach can deal with any number of speakers, here, $J = 2$ is fixed for conciseness. To record relatively diffuse noise, we let four loudspeakers simultaneously reproduce the pub noise of [21]. Results are aggregated from 25 experiments for all combinations of room, source-array distance, and SNR $\in \{10, 30\}$ dB.

No knowledge of the DOAs is assumed in the evaluation of the *weakly guided* approach. The first look direction $\varphi_1^* \in \Phi$ is selected at random. In later iterations, we set $\varphi_{i+1}^* = \varphi_i^* + \frac{360°}{J} \mod 360°$. For comparison, we also trained models for the following variants:

- Unguided: omit the branch of Fig. 4 that corresponds to $\boldsymbol{\mathcal{G}}_i$. The speaker permutation with the lowest total loss after J iterations is used to update model parameters during training (PIT).
- Strong *oracle* DOA guidance: use ground truth DOAs during both training and inference. We can consider this as an upper bound for the achievable performance.

Audio files for two examples (one using the described setup and one with $J = 3$ speakers) can be found at `https://aspire.ugent.be/demos/IWAENC2024ABsta/`.

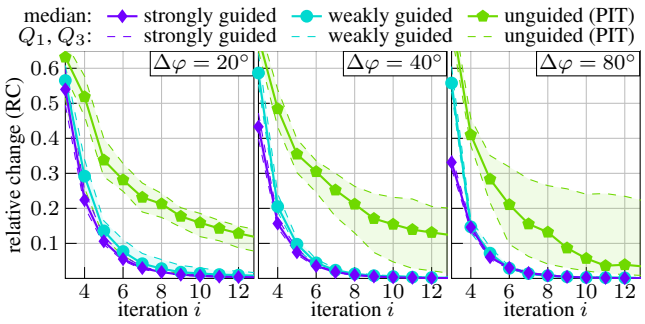


Fig. 6: Relative change $\|\mathrm{vec}\{\boldsymbol{\mathcal{F}}_i - \boldsymbol{\mathcal{F}}_{i-J}\}\|_{\ell_2} / \|\mathrm{vec}\{\boldsymbol{\mathcal{F}}_i\}\|_{\ell_2}$. The outputs of the PIT-based system often still change after many iterations. DOA guidance speeds up convergence significantly.

5.3. Convergence analysis

As the purpose of iterations $i > J$ is to refine the initial results, we expect convergence with increasing i. To determine how many iterations are needed, Fig. 6 examines the relative change (RC) in $\boldsymbol{\mathcal{F}}_i$ compared to the last iteration for the *same target speaker*. The difference between the two speakers' azimuth angles $\Delta\varphi \in \{20°, 40°, 80°\}$ varies from left to right. The median of all 200 experiments (150 for $\Delta\varphi = 80°$, as a distance of 2 m is not possible for some angles) is indicated along with the first and third quartiles (Q_1 and Q_3).

In some cases, the output of the unguided model (-●-) continues to change after many iterations. Although $\Delta\varphi = 80°$ should allow for an effective separation, the RC upper quartile remains above 0.2 after 12 iterations. Inspection of the audio files reveals that this is because there is no clearly defined target speaker in the first iteration: the degree to which either speaker is suppressed changes over time. The accordingly diminished quality of the interference and target embeddings also affects later iterations, resulting in slow convergence.

DOA guidance can resolve the ambiguity. Even with random look directions (-●-), we observe rapid convergence. By iteration $i = 5$, already, the RC median has dropped to 0.14 for the difficult case of $\Delta\varphi = 20°$, indicating that 4 iterations (2 per speaker) usually suffice. With strong DOA guidance (-◆-), convergence is slightly quicker still.

5.4. Numerical results

Fig. 7 presents scores up to iteration $i = 6$, where $i = 0$ is the input mixture. Speech intelligibility is measured with extended STOI [22] and quality with the personalized DNSMOS P.835 [23] overall score.

Given the results of Sec. 5.3, it is not surprising that the scores are significantly worse with PIT than with DOA guidance. Especially for $\Delta\varphi = 20°$, both speakers are often still present in the initial output ($i = 1$) and thus suppressed or heavily distorted in iteration $i = 2$. For this reason, the performance even deteriorates from $i = 1$ to $i = 2$ (STOI: drop from 0.53 to 0.51).

In contrast, both DOA guided variants achieve higher scores in $i = 2$ (where interference information is first available) than in $i = 1$. The best results are then obtained in later iterations ($i > 2$), when the interference *and* target embeddings (6) are nonzero. Although the weakly DOA guided model is outperformed by the oracle guidance, the two come *fairly* close at $\Delta\varphi = 80°$ (after convergence: 0.86 compared to 0.83 for STOI, 3.5 compared to 3.2 for DNSMOS). The gap increases with decreasing $\Delta\varphi$, where it becomes more difficult to resolve the permutation ambiguity based on the DOAs (for $\Delta\varphi = 20°$: 0.79 compared to 0.71 for STOI, 3.2 compared to 2.7 for DNSMOS).

Whereas only the approach of Sec. 4 was evaluated here, it is worth noting that configurations related to various prior works are obtained as special cases, thereby still allowing a comparison. For example, the strong guidance with exactly $i = 1$ iteration represents a single-step speaker extraction, where a one-hot vector encodes a target DOA, similar to [8]. The PIT model can be seen as a variant of the original recurrent selective hearing network [10] while $i \leq J$.

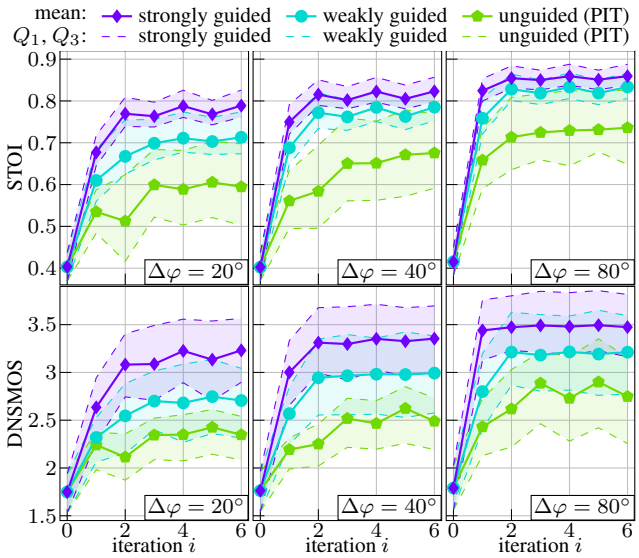


Fig. 7: Scores for the separation of $J = 2$ speakers. The one speaker is the target when i is even, the other when i is odd. Performance improves with DOA guidance and with increasing i, up to $i = 4$, confirming that the three types of prior information are used effectively.

5.5. Application to moving speaker recordings

Finally, we test the approach on a mixture of two separately recorded moving talkers, which is a scenario unseen during training. Audio files are available at `https://aspire.ugent.be/demos/IWAENC2024ABmov/`. The choice of the (time-invariant) look directions for the weak guidance was made according to the setup of Sec. 5.2. As the speakers move, it can change who is closest to a particular look direction, which causes target and interferer to swap roles. In later iterations, such switches can gradually shift backward in time, as the model only uses context from previous frames. Aside from this, we find that the weakly guided approach separates the speakers quite well. The improvement compared to PIT is again clearly apparent.

When the (time-variant) oracle DOAs are used for a strong guidance, the same speaker is extracted at all times. Only around the intersection of the two movement trajectories, the interferer is not suppressed. Although the true DOAs are unavailable in practice, separate outputs for each speaker could still be obtained by combining weak guidance with a post-processing step, opting for a strong guidance with estimated DOAs, or by appropriately adapting the training setup. This may be further investigated in the future.

6. CONCLUSIONS

Speaker separation can be performed over the course of several forward passes of the same DNN by estimating one speech signal in each step. Outputs of earlier iterations characterize target and interference. In this paper, these were jointly encoded to supply features as input that enable a gradually improving separation until convergence. Additionally, we proposed to determine in which order speakers are extracted based on their DOAs, thereby addressing the permutation ambiguity. Experiments show that even a weak guidance with random look directions improves the results considerably compared to an unguided variant. Nevertheless, a strong oracle guidance remains more reliable especially when speakers are close to each other or when the locations are time-variant. A promising extension could therefore entail integrating DOA estimation to adapt the look directions.

7. REFERENCES

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

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

[3] H. Taherian, K. Tan, and D. Wang, "Multi-channel talker-independent speaker separation through location-based training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2791–2800, 2022.

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

[5] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Spatially selective speaker separation using a DNN with a location dependent feature extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 930–945, 2024.

[6] 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 Language Technology Workshop (SLT)*, 2018, pp. 558–565.

[7] 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 International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2022, pp. 1–8.

[8] K. Tesch and T. Gerkmann, "Multi-channel speech separation using spatially selective deep non-linear filters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 542–553, 2024.

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

[10] K. Kinoshita, L. Drude, M. Delcroix, and T. Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5064–5068.

[11] N. Takahashi, S. Parthasaarathy, N. Goswami, and Y. Mitsufuji, "Recursive speech separation for unknown number of speakers," in *Proc. 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 1348–1352.

[12] Y. Zhang, Z. Chen, J. Wu, T. Yoshioka, P. Wang, Z. Meng, and J. Li, "Continuous speech separation with recurrent selective attention network," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6017–6021.

[13] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Insights into magnitude and phase estimation by masking and mapping in DNN-based multichannel speaker separation," in *Proc. Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2024, pp. 500–504.

[14] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 656–660.

[15] 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*, 2018, pp. 3942–3951.

[16] E. A. P. Habets, "RIR generator," `https://github.com/ehabets/RIR-Generator`, Accessed: December 13, 2023.

[17] 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.

[18] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proc. 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011, pp. 1509–1512.

[19] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

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

[21] 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.

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

[23] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 886–890.