Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Steering Deep Non-Linear Spatially Selective Filters for Weakly Guided Extraction of Moving Speakers in Dynamic Scenarios

Jakob Kienegger, Timo Gerkmann

Signal Processing (SP), University of Hamburg, Germany
jakob.kienegger@uni-hamburg.de, timo.gerkmann@uni-hamburg.de

Abstract

Recent speaker extraction methods using deep non-linear spatial filtering perform exceptionally well when the target direction is known and stationary. However, spatially dynamic scenarios are considerably more challenging due to time-varying spatial features and arising ambiguities, e.g. when moving speakers cross. While in a static scenario it may be easy for a user to point to the target's direction, manually tracking a moving speaker is impractical. Instead of relying on accurate time-dependent directional cues, which we refer to as strong guidance, in this paper we propose a weakly guided extraction method solely depending on the target's initial position to cope with spatial dynamic scenarios. By incorporating our own deep tracking algorithm and developing a joint training strategy on a synthetic dataset, we demonstrate the proficiency of our approach in resolving spatial ambiguities and even outperform a mismatched, but strongly guided extraction method.

Index Terms: multi-channel, moving source, spatially selective filter, target speaker extraction, direction of arrival estimation

1. Introduction

Given a real world recording containing a speech signal, the field of speech enhancement focuses on improving its quality and intelligibility by suppressing noise and reverberation artifacts. Noise sources in form of additional interfering speech signals, such as in the so-called cocktail-party problem [1], are especially challenging due to their non-stationary temporal-spectral properties. Furthermore, such scenarios require additional information about the target speaker in order to resolve the ambiguity regarding which speaker to enhance and which speakers to suppress. This particular problem setting in the realm of speech enhancement is referred to as target speaker extraction (TSE) and defined by the task of identifying, disentangling and enhancing a target speech signal from a multi-speaker recording based on a speaker-specific cue. While a variety of cue modalities have been proposed for this purpose, see e.g. [2] for an overview, spatial information becomes especially popular in the case of a stationary speaker and the availability of recordings from a microphone array. Commonly under the assumption that contributing speech sources are spatially distinguishable by their azimuth orientation to the array, a spatially selective filter (SSF) can be employed and steered in the target's direction to extract the desired speech signal [3, 4, 5]. Recently proposed data-driven implementations have gained a great deal of attention by demonstrating very high spatial selectivity with only a limited amount of microphones [3, 4, 6, 7]. By joint exploitation of spatial, temporal and spectral information, such deep non-linear SSFs drastically improve the enhancement performance over traditional linear filtering approaches as e.g. the minimum-variance distortionless response (MVDR) beamformer [8]. However, due to the resulting high spatial selectivity, they also result in an increased dependency on the provision of accurate directional information [3, 6].

A recording setup such as a seated conference meeting with a centered microphone array [9] can legitimate the assumption of stationary and directionally distinct speaker locations. The training procedure can even be adapted to increase the robustness of SSFs against small perturbations such as minor cue inaccuracies or head movement of a speaker [3]. However, in more general settings such as e.g. the dinner party scenario considered in [10], participants can move freely within the recording environment and the prior assumptions are no longer valid. With the majority of data-driven SSFs utilizing temporal context to extract the target's speech signal [3, 4, 5, 6], the change from a stationary to a dynamic acoustic scenario bears a significant domain shift. To counteract this effect, recent works include non-stationary speaker locations during training by abruptly changing their positions over the duration of the recording [5, 11]. Although spatialization with such piecewise constant trajectories is far less computationally expensive than simulating continuous motion, their discontinuity leads to a poor generalization in real recordings as demonstrated in [12] for the task of sound source localization (SSL). The arising necessity of time-dependent directional cues bears an additional challenge, since they are usually not available in practical applications. While [3, 5] have used an upstream SSL to estimate the speaker locations for the task of speech separation, oracle information is necessary to disentangle speaker trajectories in case two speakers pass each other [5]. Furthermore, during such indeterminate speaker arrangements a SSF cannot differentiate the overlapping speech signals solely based on their relative azimuth orientations toward the microphone array, see the visualization in [5]. Whereas enriching the cue with additional positional information [6, 7] or changing its modality to include the target's temporal-spectral characteristics e.g. by utilizing an enrollment utterance [11, 13, 14] can help to resolve the ambiguity, it greatly limits the applicability in real-world scenarios.

Instead of relying on accurate time-dependent directional cues, which we refer to as *strong* guidance, we propose a *weakly* guided TSE method which solely depends on the speaker's initial position. By creating a synthetic dataset modeling continuous movement, we demonstrate the significant performance improvement of SSFs in dynamic scenarios and showcase their capability to resolve spatial ambiguities by learning to differentiate temporal-spectral patterns. Finally, by introducing a *joint* training strategy, we can not only report strong results combined with our own deep tracking neural network (NN), but also paired with a very simple and inaccurate tracking algorithm.

10.21437/Interspeech.2025-1515

2. Problem definition

2.1. Target speaker extraction

Let the multi-channel observation $\mathbf{Y}$ be modeled as additive mixture of reverberant target speech $\mathbf{X}$ with noise $\mathbf{V}$. The latter contains interfering speech sources as well as additive environmental and measurement noise. In the short-time Fourier transform (STFT) domain, this acoustic setup is resembled by

$$\mathbf{Y}_{tk} = \mathbf{X}_{tk} + \mathbf{V}_{tk}\,, \tag{1}$$

with t and k denoting frame and frequency bins respectively and vectorization conducted over the channel dimension. In this work, we define the task of TSE as the recovery of the direct path propagated speech signal S_{tk}, also referred to as *dry* speech signal [3], at a predefined reference microphone. While a vast amount of methods have been proposed for this task, we will focus on deep NN algorithms computing a complex-valued mask $\mathcal{M}_{tk}$ to extract the target speech signal $\widehat{S}_{tk}$ by multiplication with a reference channel of the input Y^0_{tk}

$$\widehat{S}_{tk} = \mathcal{M}_{tk} Y^0_{tk}\,. \tag{2}$$

The data-driven estimation of time-frequency mask $\mathcal{M}_{tk}$ conditioned on a target specific cue is the center of this work.

2.2. Spatially selective filter

Spatially selective filters (SSFs) exploit positional cues to distinguish the target's speech signal from interfering sources. Since accurate positional information in three-dimensional space is usually not available, only the relative azimuth orientation of the target speaker towards the microphone array is commonly utilized for this purpose [3, 4, 5]. To some degree this can be legitimized in case the inter-microphone distance is negligible towards the distance between target speaker and recording array. On the one hand the validity of the far-field approximation [8] increases, reducing the influence of distance on the signal characteristics. Furthermore, a change in speaker height only results in a slight change in elevation angle, which leaves the azimuth direction as the most distinct spatial feature. However, there have been extensions proposed to incorporate information about elevation as well as distance into the extraction process [6, 7]. In this work we abide by common convention and constrain the cue on the azimuth angle, which we will refer to as direction of arrival (DoA) in the following. Thus, the task of the SSF lies in computing the time-frequency mask $\mathcal{M}_{tk}$ in (2) from the multi-channel input $\mathbf{Y}_{tk}$ conditioned on the DoA θ. In most realizations, the latter is provided as a one-hot encoded vector $\mathbf{z}$ corresponding to a specific spatial discretization in Θ equidistant angular regions [3, 5, 6]. While in the majority of prior work this cue is time-independent and a stationary setting is assumed [3, 4, 5, 6], we explicitly include motion to target and interfering speakers, thus, necessitating time-dependent directional information θ_t, as indicated by the index t.

3. Proposed method

3.1. Weakly guided target speaker extraction

Instead of relying on accurate *continuous* prior directional information, we relax this constraint and only assume knowledge about the *initial* starting direction θ_0 of the target speaker. In order to retain compatibility with conventional SSFs like [3, 4, 5, 6], we propose to incorporate an upstream tracking algorithm to estimate the evolution of the target's DoA θ_t based on θ_0. In analogy to TSE, we will refer to this task as target speaker tracking (TST) and avoid the term target speaker localization, as it is commonly paired with audio cues [13, 15, 16]. Figure 1 displays the resulting TSE pipeline in form of a concatenation of an upstream TST and downstream SSF algorithm.

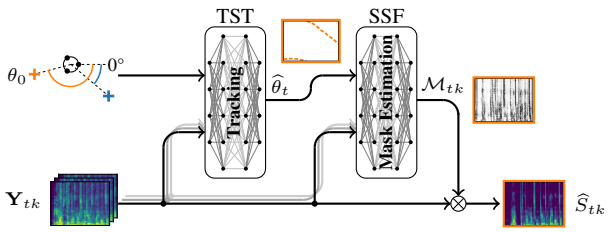


Figure 1: *Proposed weakly guided target speaker extraction (TSE) pipeline. In contrast to strongly guided TSE, our method only requires the initial DoA θ_0 to extract the speech signal $\widehat{S}_{tk}$.*

3.2. Target speaker tracking

The general problem of estimating and tracking the direction of a moving source is a well-established field of research. Classical model-based approaches such as Kalman filter (KF) [18, 19] or particle filter (PF) [20, 21] build on an underlying statistical framework and provide robust and computationally efficient algorithms. However, the performance of such algorithms substantially decreases in challenging noisy and reverberant scenarios [22]. While improvements are possible by incorporation of data-driven methods [23, 24], in case enough training data is available, recent trends are moving toward fully neural approaches. Especially deep recurrent architectures show great promise in the related field of SSL by learning intricate dependencies between observation and the temporal evolution of the DoA [17, 25, 26]. In this work we use the established causal SSL model proposed by Bohlender et al. [17] as starting point and modify it for the purpose of TST. Due to a convolutional spatial preprocessing followed by a unidirectional LSTM layer, it is commonly referred to as CNN/LSTM [5, 17]. In its original implementation the DoA estimation is encoded in a classification problem in fixed angular steps, which is a common approach for data-driven SSL algorithms [3, 27, 28]. While the resulting spatial quantization can lead to estimation inaccuracies, in our specific use-case the output modality perfectly matches the one-hot encoding $\mathbf{z}_t$ of the downstream SSF.

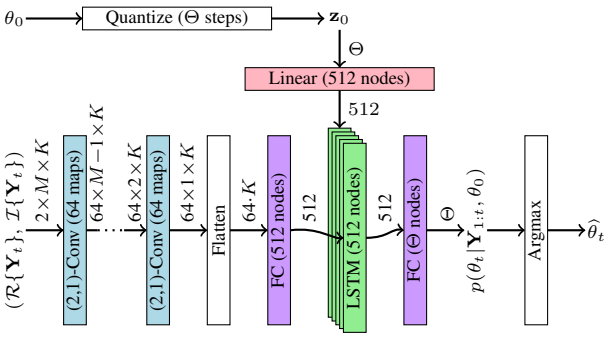


Figure 2: *Proposed deep target speaker tracking (TST) conditioned on the initial DoA θ_0. The network architecture is based on [3] and [17]. Colored layers indicate learnable parameters.*

To modify the architecture towards tracking a specific target speaker, we change the original output activation from Sigmoid to Softmax [27], making it interpretable as posterior probability $p(\theta_t|\mathbf{Y}_{1:t}, \theta_0)$. We then estimate the target's DoA $\widehat{\theta}_t$ according to the maximum a posteriori (MAP) decision rule

$$\widehat{\theta}_t = \underset{\theta_t}{\operatorname{argmax}}\, p(\theta_t|\mathbf{Y}_{1:t}, \theta_0)\,. \tag{3}$$

In order to inform the NN about the initial DoA θ_0, we choose to adopt the initialization technique employed in the FT-JNF architecture [3]. Specifically, we use the one-hot encoded DoA representation $\mathbf{z}_t$ transformed by a linear layer directly as initialization of the LSTM in the TST network. Finally, we replace the original input features in [17] with the complex multi-channel spectrogram, matching both FT-JNF and SSL proposed in [3]. The resulting deep TST architecture is illustrated in fig. 2.

4. Experiments

4.1. Dataset

For training and evaluation we create a synthetic dataset with reverberant two-speaker mixtures. Specifically, we use utterances from the LibriSpeech [29] corpus, convolve them with simulated room impulse responses (RIRs) based on the image method [30] and then combine them according to the recipe in Libri2Mix [31]. Since no additional noise is present in the mixture, the interfering signal $\mathbf{V}_{tk}$ in (1) consists of the early and late reflections of the target speech signal and the reverberant interfering speech. The acoustic room setup closely follows the work from Tesch et al. [3]. Hence, we choose a circular array of $10\,\mathrm{cm}$ diameter as recording device, which consists of three uniformly spaced omnidirectional microphones. We adhere to the reverberation times between $0.2\,\mathrm{s}$ and $0.5\,\mathrm{s}$ as well as the same simulation setup for the shoe-box rooms. At the start of each mixture, target and interfering speaker are separated by at least 10° and at a distance between 0.8 to $1.2\,\mathrm{m}$ from the array. The distance from the array as well as the height of the speaker is kept constant over the duration of each mixture while the azimuth direction may vary. For this purpose we employ a constant velocity motion model [32] defined as

$$\begin{bmatrix}\theta_{t+1}\\ \dot{\theta}_{t+1}\end{bmatrix} = \begin{bmatrix}1 & \Delta T\\ 0 & 1\end{bmatrix}\begin{bmatrix}\theta_t\\ \dot{\theta}_t\end{bmatrix} + \begin{bmatrix}\Delta T^2/2\\ \Delta T\end{bmatrix}\nu_t\,, \tag{4}$$

with the perturbation ν_t being a white, zero mean Gaussian process of variance σ^2 and implicitly taking the wrapping of θ_t into account. Since there is no constraint on the azimuth movement, intersecting trajectories of target and interfering speakers are allowed and depending on the perturbation ν_t, more or less likely to happen. To obtain a physically meaningful parametrization, we derive the distribution of the angular displacement $\Delta\theta_t$ between start and current frame t. Since this quantity is not affected by wrapping effects, sequentially evaluating the above linear state-space model leads to the Gaussian distribution

$$\Delta\theta_t \sim \mathcal{N}\left(0, \frac{\Delta T^4}{12}\left(4t^3 - t\right)\sigma^2\right)\,. \tag{5}$$

Due to its symmetry, the expected absolute displacement $\mathbb{E}\{|\Delta\theta_t|\}$ can be expressed through the closed-form solution

$$\mathbb{E}\{|\Delta\theta_t|\} = \frac{\Delta T^2}{\sqrt{6\pi}}\sqrt{4t^3 - t}\,\sigma\,. \tag{6}$$

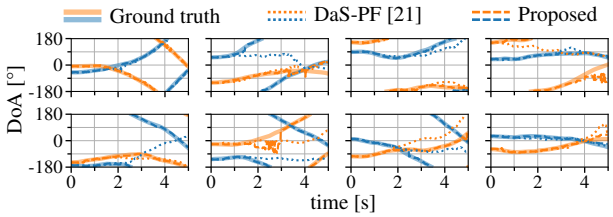


Figure 3: *DoA estimation with target speaker tracking (TST) methods presented in section 4.2. Selected trajectories correspond to* (4) *with an expected displacement* $\mathbb{E}\{|\Delta\theta_t|\}$ *of* $\frac{180^\circ}{5\mathrm{s}}$.

In the following, we will determine the standard deviation σ of the motion model's perturbation ν_t in (4) from a given mean absolute deviation $\mathbb{E}\{|\Delta\theta_t|\}$ after t frames using (6). The trajectory time step ΔT of the RIR simulations is aligned with the STFT parametrization, for which we use a square-root Hann window [33] of length $32\,\mathrm{ms}$ and $16\,\mathrm{ms}$ hop-size. Figure 3 illustrates the resulting motion dynamics corresponding to an expected angular displacement $\mathbb{E}\{|\Delta\theta_t|\}$ of $\frac{180^\circ}{5\mathrm{s}}$.

4.2. Network architectures and training details

Spatially selective filter To test our proposed methods with a strong SSF baseline we choose FT-JNF proposed by Tesch et al. in [3]. Due to its exceptional spatial filtering capabilities paired with a conceptually simple and lightweight architecture, it has seen great popularity in a variety of applications [34, 35, 36]. In order to retain causality, we slightly deviate from [3] and replace the last B-LSTM layer by an unidirectional LSTM of twice its initial parameters to roughly preserve the intended model size. We train FT-JNF with an initial learning rate of 10^{-3} and exponential decay of 10^{-2} according to the joint time and frequency domain loss function in [3].

Target speaker tracking To cope with the challenging motion dynamics in our dataset, we utilize our proposed architecture in section 3.2 as upstream TST method with the parametrization shown in fig. 2. We train the NN to minimize the cross-entropy towards the one-hot encoded ground truth DoA $\mathbf{z}_t$ with an initial learning rate of 10^{-4} and exponential decay of 10^{-2}. To give a comparison with a traditional approach, we choose a PF due to its popularity and algorithmic simplicity. Specifically, we employ the algorithm in [21] using the output power of a delay-and-sum (DaS) beamformer as likelihood, which we will refer to as DaS-PF in the following. To strengthen this baseline, we employ the motion model in (4) and match the perturbation's standard deviation σ to the dataset.

Training All NNs are trained on mixtures of $5\,\mathrm{s}$ length while resampling the randomized acoustic setup from section 4.1 including the speaker trajectories in each step. The latter becomes feasible due to the GPU accelerated implementation of the image method in [37]. We use spatial discretization of 2° which results in a total number of 180 spatial regions Θ. We employ two different training strategies for our proposed TSE pipeline in fig. 1. In the first we train the NNs independently for a total of 100 and 50 epochs regarding the SSF and TST architectures respectively. The second strategy is to train the pipeline in an end-to-end fashion, consequently, we drop the argmax operation in (3) to retain differentiability. For comparability, we pre-train the SSF and TST NNs independently for 75 and 25 epochs and then continue for the remaining 25 epochs on their concatenation regarding the loss of the FT-JNF.

Table 1: *Performance of target speaker extraction (TSE) pipelines with FT-JNF [3] and varying tracking and training strategies for expected DoA shifts* $\mathbb{E}\{|\Delta\theta_t|\}$ *of* $\frac{0°}{5s}$ / $\frac{180°}{5s}$ / $\frac{360°}{5s}$.

ID	Pipeline: Tracking	Pipeline: Training	Performance [$\frac{0°}{5s}$ / $\frac{180°}{5s}$ / $\frac{360°}{5s}$]: SI-SDR [dB] ↑	PESQ [-0.5−4.5] ↑	ESTOI [%] ↑
(0)	–	–	-6.81/-6.82/-6.83	1.13/1.13/1.12	43.9/43.2/42.9
(1)	Oracle	Static	**5.41**/2.18/1.72	**1.99**/1.61/1.56	**75.8**/69.3/67.9
(2)	Oracle	Dyn.(ours)	4.71/**4.27**/**4.02**	1.93/**1.84**/**1.80**	74.3/**72.6**/**71.8**
(3)	DaS-PF[21]	Static	**5.41**/-1.84/-4.38	**1.99**/1.33/1.20	**75.8**/52.5/40.1
(4)	DaS-PF[21]	Dyn.(ours)	4.71/-1.69/-4.20	1.93/1.28/1.21	74.3/49.0/44.1
(5)	DaS-PF[21]	Joint(ours)	4.27/**1.97**/**0.39**	1.86/**1.54**/**1.42**	72.6/**64.1**/**59.1**
(6)	Proposed	Static	**5.21**/0.73/-0.21	**1.98**/1.56/1.49	**75.5**/67.5/64.5
(7)	Proposed	Dyn.(ours)	4.41/2.37/1.38	1.93/1.75/1.65	74.0/70.2/**67.2**
(8)	Proposed	Joint(ours)	4.56/**3.44**/**2.21**	1.94/**1.77**/**1.66**	74.0/**70.3**/66.5

5. Results

We evaluate the TSE performance w.r.t. distortions, perceptual quality and intelligibility via the metrics SI-SDR [38], PESQ [39] and ESTOI [40] respectively. Table 1 displays the results using FT-JNF with different tracking and training strategies for three levels of speaker dynamics, complemented by fig. 4 regarding SI-SDR improvement with a refined resolution.

Strongly guided speaker extraction To investigate the impact of moving speakers on the extraction capabilities of a SSF, we supply FT-JNF with continuous ground truth DoA information. When training in a stationary setting, FT-JNF displays a sharp performance decline across all metrics with increasing speaker motion (1). The same network but trained on a dynamic setup, in which we uniformly draw the expected azimuth shift $\mathbb{E}\{|\Delta\theta_t|\}$ from the evaluation interval, is only slightly affected by the speaker's motion and results in improved metrics for all non-stationary cases (2). Since FT-JNF optimized in a motionless environment (1) can fully rely on distinct and stationary spatial features, the network is bound to overfit on this specific setup achieving a very high spatial selectivity [3]. However, when dropping this constraint, the SSF is not able to cope with the increasingly challenging acoustic characteristics as well as incapable of handling overlapping speakers due to solely relying on the provided spatial cue. On the other hand, as shown in fig. 6, training the identical network in a dynamic setting, FT-JNF is able to resolve the ambiguity and attenuate the interfering speaker sufficiently. This demonstrates that although being conditioned on a *spatial* cue, SSFs are also capable of distinguishing and leveraging *temporal-spectral* characteristics.

Weakly guided speaker extraction Instead of continuous oracle DoA information, our proposed *weakly* guided TSE pipeline only relies on the target's initial DoA by utilizing an upstream TST system. To compare both tracking algorithms discussed in this work, we visualize sample trajectories in fig. 3 and report frame-wise accuracy (ACC) with a 5° margin [15, 16] and angular error (AE) in fig. 5. Although the DaS-PF utilizes the motion model with oracle parametrization, which explains the perfect scores in the stationary case, the performance decreases rapidly with increasing speaker movement. While our proposed network shows a much stronger and consistent overall performance, it also follows the same trend. For this reason the simple concatenation of SSF and TST methods leads to a significant performance decrease in all cases $(3, 4, 6, 7)$. However, with our proposed *joint* training strategy, we are able to counteract this effect and even surpass the strongly guided TSE configuration in (1) regarding distortion and speech quality metrics for the non-stationary setups, underlining the importance of a dynamic dataset during training. Figure 6 illustrates the strong influence of erroneous DoA estimations from our TST algorithm on FT-JNF (6, 7), which only the joint training (8) is able to resolve. The biggest gain can be observed for the DaS-PF (5). Although providing highly inaccurate DoA estimates, the FT-JNF is able to adapt with our joint training strategy and greatly improve across all metrics, see also fig. 4.

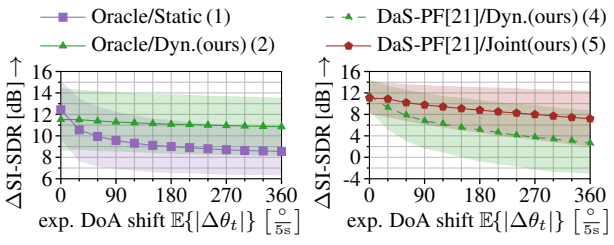


Figure 4: *Selected TSE configurations evaluated with a metric sensitive to distortions. Shaded areas indicate std. deviation.*

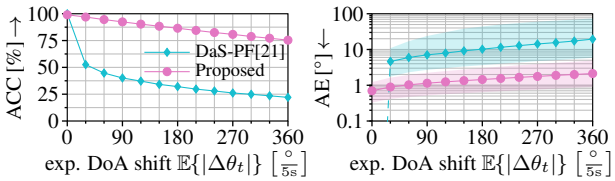


Figure 5: *Accuracy (ACC) and median angular error (AE) of TST algorithms. Shaded areas indicate 25% and 75% quartiles.*

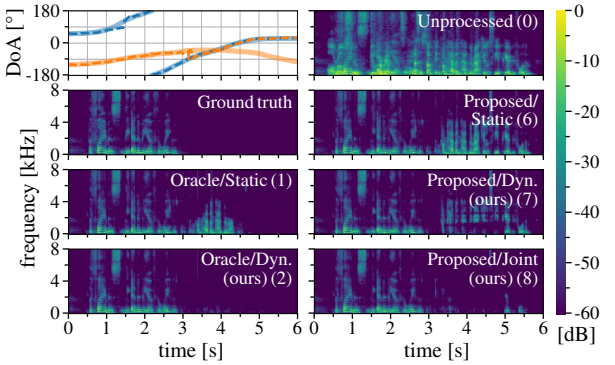


Figure 6: *Strongly (1, 2) and weakly (6, 7, 8) guided extraction results with FT-JNF [3] and varying tracking/training configurations during crossing of target (•) and interfering speaker (•).*

6. Conclusion

In this work we proposed a *weakly* guided TSE pipeline which solely requires information about the target's initial direction. By developing a synthetic dataset with continuous speaker movement, we could not only demonstrate the significant performance improvement of a commonly used SSF in dynamic environments, but also its capability to resolve spatial ambiguities by learning to differentiate temporal-spectral patterns. Finally, by introducing a *joint* training strategy, we could surpass the performance of a mismatched, but strongly guided system with our own deep tracking NN and also illustrate how SSFs can leverage highly inaccurate directional cues and achieve competitive performance even with a very simple tracking algorithm.

7. Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant 508337379. Computational resources were provided by Regional Computer Center (RRZ) of the University of Hamburg (UHH) and Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project f104ac. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU and RRZ hardware are partially funded by the DFG under grants 440719683 and 498394658 respectively.

8. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *J. Acoust. Soc. Am.*, vol. 25, 1953.

[2] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Proc. Magazine*, vol. 40, 2023.

[3] K. Tesch and T. Gerkmann, "Multi-channel speech separation using spatially selective deep non-linear filters," *IEEE/ACM TASLP*, vol. 32, 2024.

[4] A. Briegleb, M. M. Halimeh, and W. Kellermann, "Exploiting spatial information with the informed complex-valued spatial autoencoder for target speaker extraction," in *IEEE ICASSP*, 2023.

[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 TASLP*, vol. 32, 2024.

[6] A. Pandey, S. Lee, J. Azcarreta, D. Wong, and B. Xu, "All neural low-latency directional speech extraction," in *Interspeech*, 2024.

[7] R. Gu and Y. Luo, "ReZero: Region-customizable sound extraction," *IEEE/ACM TASLP*, 2024.

[8] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Hoboken, NY, USA: Wiley, 2006.

[9] 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 *IEEE ICASSP*, 2020.

[10] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines," in *Interspeech*, 2018.

[11] S. He, J. Liu, H. Li, Y. Yang, F. Chen, and X. Zhang, "3S-TSE: Efficient three-stage target speaker extraction for real-time and low-resource applications," in *IEEE ICASSP*, 2024.

[12] A. Bohlender, L. Roelens, and N. Madhu, "Improved deep speaker localization and tracking: Revised training paradigm and controlled latency," in *IEEE ICASSP*, 2023.

[13] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "L-SpEx: Localized target speaker extraction," in *IEEE ICASSP*, 2022.

[14] G. Li, S. Liang, S. Nie, W. Liu, M. Yu, L. Chen, S. Peng, and C. Li, "Direction-aware speaker beam for multi-channel speaker extraction." in *Interspeech*, 2019.

[15] G. Li, W. Xue, W. Liu, J. Yi, and J. Tao, "GCC-Speaker: Target speaker localization with optimal speaker-dependent weighting in multi-speaker scenarios," in *IEEE ICASSP*, 2023.

[16] Y. Chen, X. Qian, Z. Pan, K. Chen, and H. Li, "LocSelect: Target speaker localization with an auditory selective hearing mechanism," in *IEEE ICASSP*, 2024.

[17] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Exploiting temporal context in CNN based multisource DoA estimation," *IEEE/ACM TASLP*, vol. 29, 2021.

[18] J. Traa and P. Smaragdis, "A wrapped kalman filter for azimuthal speaker tracking," *IEEE Signal Proc. Letters*, vol. 20, no. 12, 2013.

[19] M. Murase, S. Yamamoto, J.-M. Valin, K. Nakadai, K. Yamada, K. Komatani, T. Ogata, and H. Okuno, "Multiple moving speaker tracking by microphone array on mobile robot," in *Interspeech*, 2005.

[20] F. Dong, L. Xu, and X. Li, "Particle filter algorithm for DoA tracking using co-prime array," *IEEE Comm. Letters*, vol. 24, 2020.

[21] D. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, 2003.

[22] Q. Zhang, Z. Chen, and F. Yin, "Speaker tracking based on distributed particle filter in distributed microphone networks," *IEEE Trans. on Systems, Man, and Cybern.: Systems*, vol. 47, 2017.

[23] G. Revach, N. Shlezinger, X. Ni, A. L. Escoriza, R. J. G. van Sloun, and Y. C. Eldar, "KalmanNet: Neural network aided kalman filtering for partially known dynamics," *IEEE Trans. on Signal Proc.*, 2022.

[24] X. Chen, H. Wen, and Y. Li, "Differentiable particle filters through conditional normalizing flow," in *IEEE FUSION*, 2021.

[25] B. Yang, H. Liu, and X. Li, "SRP-DNN: Learning direct-path phase difference for multiple moving sound source localization," in *IEEE ICASSP*, 2022.

[26] H. Yin, M. Ge, Y. Fu, G. Zhang, L. Wang, L. Zhang, L. Qiu, and J. Dang, "MIMO-DoAnet: Multi-channel input and multiple outputs DoA network with unknown number of sound sources," in *Interspeech*, 2022.

[27] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *IEEE ICASSP*, 2015.

[28] G. K. Papageorgiou, M. Sellathurai, and Y. C. Eldar, "Deep networks for DoA estimation in low SNR," *IEEE Trans. on Signal Proc.*, vol. 69, 2021.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE ICASSP*, 2015.

[30] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, 1979.

[31] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," 2020. [Online]. Available: https://arxiv.org/abs/2005.11262

[32] X. Rong Li and V. Jilkov, "Survey of maneuvering target tracking. Part I. Dynamic models," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 39, 2003.

[33] S. Shimauchi and H. Ohmuro, "Accurate adaptive filtering in square-root Hann windowed short-time fourier transform domain," in *IEEE ICASSP*, 2014.

[34] J. Wechsler, S. R. Chetupalli, M. M. Halimeh, O. Thiergart, and E. A. P. Habets, "Neural directional filtering: Far-field directivity control with a small microphone array," in *IWAENC*, 2024.

[35] A. Briegleb and W. Kellermann, "Spatially constrained vs. unconstrained filtering in neural spatiospectral filters for multichannel speech enhancement," in *EUSIPCO*, 2024.

[36] B. Lentz and R. Martin, "Utilizing head rotation data in DNN-based multi-channel speech enhancement for hearing aids," in *IWAENC*, 2024.

[37] D. Diaz-Guerra, A. Miguel, and J. R. Beltrán, "gpuRIR: A Python library for room impulse response simulation with GPU acceleration," *Multimedia Tools and Applications*, vol. 80, 2018.

[38] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or well done?" in *IEEE ICASSP*, 2019.

[39] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE ICASSP*, 2001.

[40] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM TASLP*, vol. 24, 2016.