ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747141

LOCATION-BASED TRAINING FOR MULTI-CHANNEL TALKER-INDEPENDENT SPEAKER SEPARATION

Hassan Taherian[1], *Ke Tan*[1], and *DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
{taherian.1, tan.650}@osu.edu, dwang@cse.ohio-state.edu

ABSTRACT

Permutation-invariant training (PIT) is a dominant approach for addressing the permutation ambiguity problem in talker-independent speaker separation. Leveraging spatial information afforded by microphone arrays, we propose a new training approach to resolving permutation ambiguities for multi-channel speaker separation. The proposed approach, named location-based training (LBT), assigns speakers on the basis of their spatial locations. This training strategy is easy to apply, and organizes speakers according to their positions in physical space. Specifically, this study investigates azimuth angles and source distances for location-based training. Evaluation results on separating two- and three-speaker mixtures show that azimuth-based training consistently outperforms PIT, and distance-based training further improves the separation performance when speaker azimuths are close. Furthermore, we dynamically select azimuth-based or distance-based training by estimating the azimuths of separated speakers, which further improves separation performance. LBT has a linear training complexity with respect to the number of speakers, as opposed to the factorial complexity of PIT. We further demonstrate the effectiveness of LBT for the separation of four and five concurrent speakers.

Index Terms— Multi-channel speaker separation, permutation invariant training, location-based training.

1. INTRODUCTION

Recent speaker separation methods based on deep neural networks (DNNs) have substantially improved separation performance [1, 2, 3]. To train a talker-independent separation model, where test speakers can be different from training ones, each output layer of a DNN model needs to be associated with one distinct speaker in the mixture [4]. Ambiguity in speaker assignment would lead to conflicting gradients during training. This permutation ambiguity problem also arises in DNN based speaker diarization [5] and multi-source speaker localization [6]. Main solutions to this problem include deep clustering [7] and permutation invariant training (PIT) [8]. In deep clustering, a DNN maps time-frequency units to embedding vectors with an objective function that is invariant to speaker permutations. These embedding vectors are then clustered via the K-means algorithm to estimate the ideal binary mask. On the other hand, PIT resolves the permutation ambiguity by examining the losses from all possible output-speaker permutations, and it does not require an additional clustering step.

Deep clustering and PIT were originally developed for monaural speaker separation. The availability of multi-channel recordings provides a spatial dimension, which is missing in monaural recordings. We believe that the permutation ambiguity problem can be naturally avoided by leveraging spatial relations of multiple speakers. It is a basic fact that multiple speakers cannot occupy the same spatial location. In this study, we propose a new training approach to achieving multi-channel talker-independent speaker separation. To resolve the permutation ambiguity problem, we propose location-based training (LBT), which assigns DNN output layers according to speaker locations. Specifically, we investigate azimuth-based and distance-based training, which makes assignments based on speaker azimuth angles and distances relative to a microphone array.

Our separation model uses multi-channel complex ratio masking (MC-CRM). Evaluation results show that azimuth-based training outperforms PIT in both anechoic and reverberant environments, while distance-based training is more superior in conditions where speakers have close azimuths. To combine the relative advantages of azimuth-based and distance-based training, we dynamically select the two training criteria on the basis of azimuth estimates of separated speakers. In this case, speaker localization is performed by mask-weighted generalized cross-correlation with phase transform (GCC-PHAT) [9].

In contrast to PIT and its variants whose training complexity are factorial or polynomial to the number of speakers [10, 11], LBT has a linear computational complexity. Given the low complexity of LBT, multi-channel DNN models can be

This research was supported in part by a National Science Foundation grant (ECCS-1808932), the Ohio Supercomputer Center, and the Pittsburgh Supercomputer Center (NSF ACI-1928147).

trained efficiently for a large number of concurrent speakers. Moreover, separated speakers are naturally ordered according to their spatial locations. This facilitates the integration of a speaker separation model with downstream speech processing tasks such as speaker localization, diarization, recognition, and automatic speech recognition [3, 12, 13].

This work expands our preliminary study [14] which illustrates the effectiveness of azimuth-based training for two-speaker mixtures in anechoic conditions. A recent study also considers ordering speakers based on azimuth angles for the task of multi-source speaker localization [6].

2. SYSTEM DESCRIPTION

2.1. Location-based training

A primary approach to talker-independent speaker separation utilizes utterance-level PIT to address the permutation ambiguity problem [2, 3, 13]. Utterance-level PIT uses fixed output-speaker pairings for a whole utterance, and selects the optimal pairing that minimizes the loss function over all possible speaker permutations [8]:

$$\mathcal{L}_{\text{PIT}} = \min_{\phi_1,\ldots,\phi_N \in \Phi} \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n, S_{\phi_n}), \tag{1}$$

where $\hat{S}$ and S are estimated and clean speech signals in the short-time Fourier transform (STFT) domain, respectively. $\mathcal{L}$ denotes a loss function, and symbol Φ is the set of all permutations of N speakers.

With the assumption that speakers are still, we propose to utilize the spatial locations of speakers to resolve the permutation ambiguity problem for multi-channel talker-independent speaker separation. In this study, we explore LBT based on speaker azimuth angles and distances relative to the center of a microphone array.

Let $\theta_1, \theta_2, \ldots \theta_N \in [0, 2\pi)$ be the sorted speaker azimuths relative to the microphone array. The loss of azimuth-based training is defined as:

$$\mathcal{L}_{\text{Azimuth}} = \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n, S_{\theta_n}). \tag{2}$$

Fig. 1 illustrates LBT with 3 speakers. In the case of azimuth-based training output-speaker assignments follow the azimuths order, where the first output is tied to the speaker with the smallest azimuth and the last output is tied to the speaker with the largest azimuth. Note that the azimuth range is dependent on the array geometry. For linear arrays, the azimuth range should be in $[0, \pi)$, due to the well-documented front-back confusion of linear arrays. Similarly, we formulate distance-based training as:

$$\mathcal{L}_{\text{Distance}} = \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n, S_{d_n}), \tag{3}$$

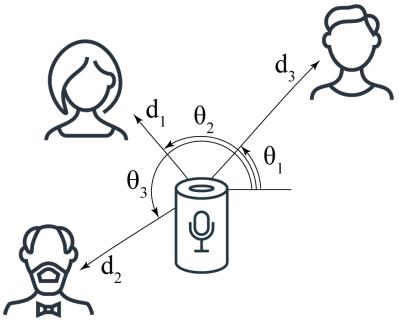

Fig. 1: Illustration of new training criteria based on speaker azimuths and distances relative to a microphone array.

where $d_1, d_2, \ldots, d_N$ are speaker distances to the microphone array in an ascending order. With this criterion, we assign the nearest speaker to the first output layer and the farthest speaker to the last output layer (see Fig. 1).

In both criteria, output-speaker assignments are based on the relative positions of speakers. By leveraging spatio-temporal patterns in the multi-channel input, LBT resolves the permutation ambiguity problem through the consistent pairings of DNN output layers and speaker locations.

2.2. Multi-channel complex ratio masking

Assuming a fixed array geometry, MC-CRM estimates the complex spectrogram of target speech received at the reference microphone from that of the multi-channel noisy mixture. It is shown that MC-CRM can implicitly learn the spectral and spatial information within the array signals [15]. We employ the Dense-UNet architecture proposed in [16] for single- and multi-channel complex ratio masking. We stack the real and imaginary components of the mixture STFT at all microphones and pass them into the Dense-UNet as input. The model estimates the cIRM (complex ideal ratio mask) [17], which is then multiplied by the complex spectrogram of the input mixture at the reference microphone.

The Dense-UNet architecture includes 4 downsampling and upsampling layers interleaved with 9 densely-connected convolutional blocks. Each dense block contains 5 convolutional layers, each of which has 64 channels, a kernel size of 3×3 and a stride of 1×1. The middle layer in each dense block is replaced with a frequency mapping layer to deal with inconsistencies between different frequency bands [16]. We adopt the loss function in [15] for each output-speaker pair, which is based on ℓ_1 norm of real and imaginary spectrograms of estimated and target speech with an additional magnitude loss term:

$$\mathcal{L}_{\text{RI+Mag}}(\hat{S}, S) = \mathcal{L}_{\text{RI}}(\hat{S}, S) + \left\| |\hat{S}| - |S| \right\|_1, \tag{4}$$

where $|S|$ and $|\hat{S}|$ represent the target and estimated magnitude spectrograms, and $|\hat{S}|$ is calculated from the estimated real and imaginary components $\hat{S}^{(r)}$ and $\hat{S}^{(i)}$:

$$|\hat{S}| = \sqrt{(\hat{S}^{(r)})^2 + (\hat{S}^{(i)})^2}. \tag{5}$$

In addition,

$$\mathcal{L}_{\text{RI}}(\hat{S}, S) = \left\| \hat{S}^{(r)} - S^{(r)} \right\|_1 + \left\| \hat{S}^{(i)} - S^{(i)} \right\|_1. \quad (6)$$

3. EXPERIMENTAL SETUP

Our experiments use simulated room impulse responses (RIRs) for evaluation. We generate RIRs for a 7-channel circular microphone array using the image method [18, 19]. The microphone array comprises 6 microphones uniformly distributed on a circle with a radius of 4.25 cm and one microphone at the center of the circle. We simulate rectangular rooms with random length, width and height dimensions in the range of [4×4×3, 6×6×4] meters, with the microphone array placed in the center of the room.

The speech sources are placed in positions randomly selected from 72 candidate azimuth positions in the range of -180° to 180° with a 5° resolution. For a speaker pair (i, j), the source-array distances d_i and d_j are randomly selected such that $|d_i - d_j| \geqslant 0.2$ m. Moreover, the minimum source-array distance is set to 0.3 m. We assume that speech sources are placed at the same height as the microphone array.

We create speech mixtures with 2 and 3 speakers in both anechoic and reverberant conditions. The multi-channel mixtures are created by spatializing the WSJ0-2mix and WSJ0-3mix datasets [7] with the simulated RIRs, which include 20000, 5000 and 3000 mixtures in the training, validation and test sets, respectively. For the reverberant mixtures, the reverberation time (T60) is randomly sampled between 0.15 and 0.6 seconds. Note that we treat the center microphone as the reference microphone. For all speakers, the direct-path (anechoic) signal at the reference microphone is used as the target signal. All signals are sampled at 16 kHz.

4. EVALUATION RESULTS

We report the results in terms of signal-to-distortion ratio (SDR) [20], scale-invariant signal-to-noise ratio (SI-SNR), perceptual evaluation of speech quality (PESQ), and extended short-time objective intelligibility (ESTOI). As a comparison baseline, we also report the results for the PIT-based single-channel CRM (SC-CRM).

Table 1 presents the MC-CRM results with different training criteria in the reverberant condition. The first two rows give results with the 5° azimuth resolution. Regardless of the training criterion, MC-CRM leads to significant performance improvement compared to SC-CRM on two-speaker and three-speaker mixtures. We observe that MC-CRM with azimuth-based training outperforms PIT in all metrics. Although distance-based training underperforms azimuth-based training, it yields comparable results to PIT.

We additionally train the MC-CRM model on reverberant two-speaker mixtures with a 1° resolution of azimuth spacing. As shown in the third row of Table 1, similar trends occur for location-based MC-CRM with this finer spatial resolution. To further investigate the effect of LBT, we evaluate MC-CRM on different sets of reverberant two-speaker mixtures where the difference between speaker azimuths is constrained. The results are shown in Fig. 2. We observe that the performance of the models with PIT and azimuth-based training significantly degrades when azimuth differences are small. Not surprisingly, distance-based training is relatively insensitive to azimuths and outperforms the other two methods.

Table 1: ESTOI (%), PESQ, SI-SNR (dB) and SDR (dB) of different training criteria on reverberant 2-speaker and 3-speaker mixtures with 5° and 1° resolutions of azimuth spacing. 'Combined' refers to the combination of azimuth-based and distance-based training.

		Criterion	ESTOI	PESQ	SI-SNR	SDR
2-speaker / 5°	Unprocessed	–	37.42	1.61	-8.15	-1.72
	SC-CRM	PIT	63.44	2.27	-0.28	3.77
	MC-CRM	PIT	78.47	2.90	5.71	8.94
	MC-CRM	Azimuth	**82.07**	**3.06**	**7.01**	**9.91**
	MC-CRM	Distance	80.01	2.96	6.45	9.07
3-speaker / 5°	Unprocessed	–	27.78	1.36	-9.49	-4.64
	SC-CRM	PIT	42.92	1.62	-3.83	0.57
	MC-CRM	PIT	67.34	2.46	4.40	6.97
	MC-CRM	Azimuth	**69.21**	**2.58**	**4.79**	**7.92**
	MC-CRM	Distance	66.63	2.41	4.24	6.62
2-speaker / 1°	Unprocessed	–	37.36	1.61	-8.15	-1.75
	MC-CRM	PIT	74.78	2.74	4.64	7.88
	MC-CRM	Azimuth	80.98	3.03	6.66	9.74
	MC-CRM	Distance	79.75	2.95	6.43	9.13
	Combined	–	**81.33**	**3.04**	**6.76**	**9.84**

To take advantage of both azimuth-based and distance-based training, we perform source localization to estimate the speaker azimuths, and use these estimates to select outputs from the better model. The azimuth of speaker k can be well estimated from a speech mixture using mask-weighted GCC-PHAT [9, 21]:

$$\underset{\tau}{\text{argmax}} \sum_{(p,q)\in\Omega} \sum_{t,f} \lambda_k \text{GCC}_{p,q}(t, f, \tau), \quad (7)$$

where $\text{GCC}_{p,q}(t, f, \cdot)$ represents the GCC-PHAT function for microphone pair (p, q) at time t and frequency f. Symbol τ denotes the time delay corresponding to a candidate azimuth, and Ω is the set of all microphones pairs. Moreover, λ_k is a ratio mask for speaker k, computed using the mixture STFT Y_{ref} at the reference microphone:

$$\lambda_k = \frac{|\hat{S}_k|^2}{|\hat{S}_k|^2 + |Y_{\text{ref}} - \hat{S}_k|^2}. \quad (8)$$

For the 1° resolution experiment, we use an empirical threshold of 20° for dynamic criterion selection. Specifically, we

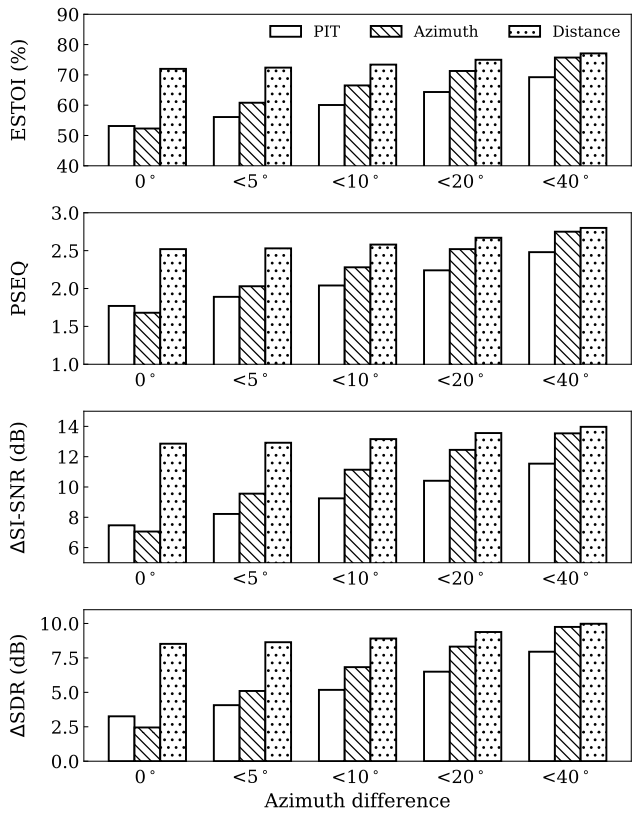

Fig. 2: ESTOI, PESQ, SI-SNR improvement (ΔSI-SNR) and SDR improvement (ΔSDR) of different training criteria with constraint azimuth difference. The models are trained on reverberant WSJ-2mix with 1° resolution.

select azimuth-based training if the computed azimuth difference is larger than 20°, and distance-based training otherwise. As shown in Table 1, such a combination further improves the results. Note that only 12% of mixtures in this test set contain speakers with an azimuth difference less than 20°.

We present the evaluation results in the anechoic condition in Table 2. Similar to Table 1, azimuth-based training achieves superior performance to PIT for two- and three-speaker mixtures. However, the performance of distance-based training significantly degrades in the anechoic condition. We conjecture that distance-based training implicitly leverages direct-to-reverberant ratios (DRRs) from different speakers for speaker separation. The DRR is inversely proportional to the square of the source-microphone distance in reverberant environments [22, 23]. As the source-microphone distance increases, the energy of the direct sound decreases while the energy of the reverberant sounds remains roughly constant. In the reverberant condition, the model trained with the distance criterion may learn to assign the speaker with the highest DRR to the first output layer and the second highest DRR to the second output layer, and so on. In an anechoic room, the DRR is infinite and thus cannot serve as a discriminative cue to separate between nearer and farther speakers. However, we note that anechoic conditions do not occur in the real world.

Table 2: Comparison of different training criteria for mixtures with the various number of speakers in the anechoic condition.

		Criterion	ESTOI	PESQ	SI-SNR	SDR
2-speaker	Unprocessed	–	56.11	1.89	0.00	0.13
	SC-CRM	PIT	83.01	2.88	11.12	11.55
	MC-CRM	PIT	97.60	4.03	24.51	25.09
	MC-CRM	Azimuth	**98.49**	**4.09**	**26.10**	**26.68**
	MC-CRM	Distance	82.47	2.97	10.77	11.30
3-speaker	Unprocessed	–	38.54	1.48	-4.43	-4.12
	SC-CRM	PIT	60.75	2.05	4.33	5.10
	MC-CRM	PIT	85.10	3.21	14.01	14.58
	MC-CRM	Azimuth	**90.82**	**3.51**	**17.13**	**17.67**
	MC-CRM	Distance	69.06	2.45	7.27	8.00
4-speaker	Unprocessed	–	29.36	1.31	-7.03	-6.51
	MC-CRM	PIT	70.61	2.64	8.19	9.04
	MC-CRM	Azimuth	**81.48**	**3.02**	**11.76**	**12.41**
5-speaker	Unprocessed	–	23.94	1.22	-8.72	-8.06
	MC-CRM	PIT	61.39	2.30	4.73	5.88
	MC-CRM	Azimuth	**70.11**	**2.56**	**7.24**	**8.07**

We have also evaluated azimuth-based training with four- and five-speaker mixtures. The same simulation procedure for the anechoic condition is used to generate a spatialized version of the WSJ0-4mix and WSJ0-5mix datasets [24]. Azimuth-based training outperforms PIT in both four- and five-speaker mixtures. In addition, as mentioned earlier, the training complexity advantage of LBT over PIT is a lot more evident for such mixtures. The results suggest that LBT can be potentially used for end-to-end diarization with a large number of speakers [5].

5. CONCLUDING REMARKS

We have proposed location-based training as a new training approach for multi-channel talker-independent speaker separation. We have developed two new training criteria based on speaker azimuth angles and distances to resolve the permutation ambiguity problem. In addition, azimuth-based and distance-based training can be combined to further improve separation performance. LBT outperforms PIT in separation performance as well as training complexity. Future work will extend LBT to separate moving speakers, and nonspeech sources.

6. REFERENCES

[1] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2840–2849, 2021.

[2] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, 2020, pp. 46–50.

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

[4] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1702–1726, 2018.

[5] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," in *Proc. Interspeech*, 2019, pp. 4300–4304.

[6] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," *arXiv:2102.07955*, 2021.

[7] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[8] 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, Lang. Process.*, vol. 25, pp. 1901–1913, 2017.

[9] Z.-Q. Wang, X. Zhang, and D. L. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 178–188, 2019.

[10] S. Dovrat, E. Nachmani, and L. Wolf, "Many-speakers single channel speech separation with optimal permutation training," *arXiv:2104.08955*, 2021.

[11] H. Tachibana, "Towards listening to 10 people simultaneously: An efficient permutation invariant training of audio source separation using sinkhorn's algorithm," in *Proc. ICASSP*. IEEE, 2021, pp. 491–495.

[12] T. Yoshioka, I. Abramovski, C. Aksoylar, Z. Chen, M. David, D. Dimitriadis, Y. Gong, I. Gurvich, X. Huang, Y. Huang *et al.*, "Advances in online audio-visual meeting transcription," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding,*, 2019, pp. 276–283.

[13] H. Taherian and D. L. Wang, "Time-domain loss modulation based on overlap ratio for monaural conversational speaker separation," in *Proc. ICASSP*, 2021, pp. 5744–5748.

[14] K. Tan, "Convolutional and recurrent neural networks for real-time speech separation in the complex domain," Ph.D. dissertation, The Ohio State University Department of Computer Science and Engineering, 2021.

[15] Z.-Q. Wang and D. L. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," in *Proc. ICASSP*, 2020, pp. 486–490.

[16] Y. Liu and D. L. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 2092–2102, 2019.

[17] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 483–492, 2016.

[18] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[19] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP*, 2018, pp. 351–355.

[20] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 1462–1469, 2006.

[21] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Process.*, vol. 24, pp. 320–327, 1976.

[22] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[23] M. Zohourian and R. Martin, "Binaural direct-to-reverberant energy ratio and speaker distance estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 92–104, 2020.

[24] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proc. ICML*, 2020, pp. 7164–7175.