A JOINT TRAINING FRAMEWORK OF MULTI-LOOK SEPARATOR AND SPEAKER EMBEDDING EXTRACTOR FOR OVERLAPPED SPEECH

Naijun Zheng[1], *Na Li*[2], *Bo Wu*[2], *Meng Yu*[2], *JianWei Yu*[1], *Chao Weng*[2], *Dan Su*[2], *XunYing Liu*[1], *Helen Meng*[1]

[1] The Chinese University of Hong Kong, [2] Tencent AI Lab

ABSTRACT

In multi-talker cases, overlapped speech degrades the speaker verification (SV) performance dramatically. To tackle this challenging problem, speech separation with multi-channel techniques can be adopted to extract each speaker's signals to improve the SV performance. In this paper, a joint training framework of the front-end multi-look speech separator and the back-end speaker embedding extractor is proposed for multi-channel overlapped speech. To better leverage the complementarity between the speech separator and the speaker embedding extractor, several training strategies are proposed to jointly optimize the two modules. Experimental results show that the proposed joint training framework significantly outperforms the individual SV system by around 52% relative EER reduction. Additionally, the robustness of the proposed framework is further evaluated under different conditions.

Index Terms— Speaker verification, multi-channel, multi-look, overlapped speech, speech separation

1. INTRODUCTION

In the last decade, rapid progress has been made in SV area with the introduction of deep neural networks (DNNs) to extract distinguishable embeddings from speaker utterances, such as d-vectors [1] and x-vectors [2]. With increasing larger datasets and deeper networks, SV research is expanding from not only clean and single-speaker conditions, but also more challenging far-field scenarios, where noise, reverberation and overlapped speech are predominant. Overlapped speech occurs frequently in multi-speaker scenarios, where the information from different speakers is mixed, leading to an undesirable degradation in many speech processing tasks, such as automatic speech recognition (ASR) [3], speaker recognition/verification and diarization.

There are mainly two ways to alleviate the deterioration caused by these factors: one way is to improve the robustness of the systems by data augmentation during training [2], which relies on the powerful capacity of the DNNs; while the other is to introduce effective front-ends to pre-process the source speech, such as the speech enhancement and separation. In the latter case, when a microphone array is available, beamforming techniques can utilize spatial information to extract the target speech signals and suppress the interference signals. There are various beamforming techniques, such as delay and sum [4], Minimum Variance Distortionless Response (MVDR) [5] methods and multi-look directions based methods [6], which have been widely applied in the pre-possessing step for ASR [7] and keyword spotting [8] tasks.

Several works have been proposed to investigate the SV systems assisted with multi-channel techniques. In [9], to reduce the reverberation effect, mask-based neural networks with several beamformers were proposed for GMM i-vector systems. In [10], a framework employing generalized eigenvalue beamformer for x-vector extractor was proposed, which found that the system with the re-trained beamformer outperformed the system with only independent training. Previous work that consider overlapped speech for the SV task are limited. A single-channel front-end extractor was proposed in [11] to extract the target speaker's speech based on the enrollment speaker information, which greatly improved the performance of the GMM i-vector system in a pipeline mode.

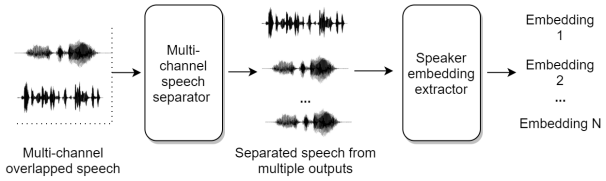


Fig. 1. The integrated system of the speech separator and the speaker embedding extractor.

However, there are few, if any, previous work on a tighter integration between a multi-channel front-end and a back-end network via partial and/or full joint training for SV tasks, especially on the overlapped speech. In this paper, we propose to jointly train a multi-look separation network with a back-end SV network and several joint training strategies are investigated. Different from prior, the proposed approach focuses on the multi-channel overlapped speech condition, where the utterances are mixed among up to two speakers. To alleviate the objective function mismatch between the separation front-ends and the SV back-end, multi-task loss functions are also explored.

The rest of the paper is organized as follows: In section 2, each module of the proposed framework is detailed. Section 3 describes our experiment setup. The performance of the joint training framework is analysed in section 4. The conclusions are drawn in section 5.

2. JOINT TRAINING FRAMEWORK

The joint training framework is shown in Figure 1, where the front-end separation network is expected to have multiple outputs and each output is corresponding to a single speaker. A back-end speaker embedding extractor is then employed to extract discriminative speaker embeddings. The embedding extractor, the separation modules, and joint training strategies are described in the following.

2.1. Speaker embedding extractor

The DNN based speaker embedding extractors have been proven to significantly outperform conventional i-vector based systems. With softmax-like objective functions [12] , the networks are usually discrimitively trained as a speaker classification [13]. A temporal av-

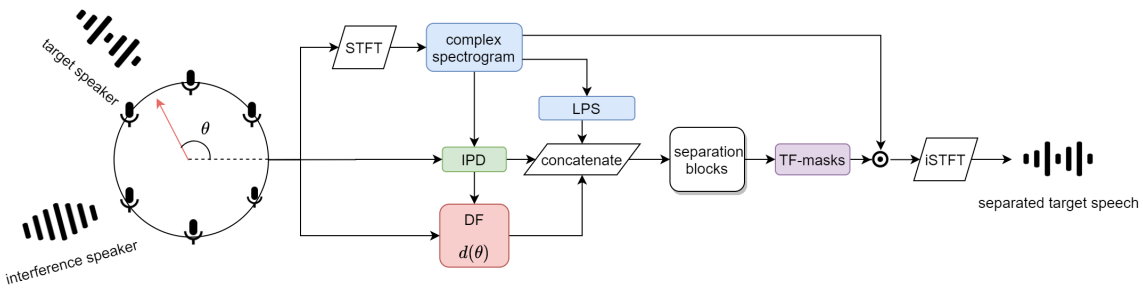


Fig. 2. Multi-channel speech separation network with look direction θ.

erage pooling (TAP) layer is used to aggregate the frame-level representation into a fixed-length utterance-level representation, which is further projected to a low-dimensional speaker embedding via a fully connection layer [12]. Given the speaker embedding extractor, cosine similarity is usually used as the back-end for evaluation.

The popular ResNet34 architecture [14] is employed as the backbone of the speaker extractor in this paper. The residual connections between the frame-level layers have been demonstrated to enhance the representation learned by the network. To further improve the discriminating power of the learned speaker embeddings, additive angular margin softmax (AAM-softmax)[15] loss function is used as the optimization objective.

2.2. Multi-channel speech separator

Figure 2 shows the diagram of the multi-channel speech separation network with a specific looking direction θ. The waveform of each channel is first converted into complex frequency-time domain using a STFT convolution encoder. Then, the logarithm power spectrum (LPS) can be obtained from the magnitude of the spectrograms. Given the microphone pair $\bar{m} = (m_1, m_2)$, we can compute the inter-channel phase difference (IPD) as follows:

$$\text{IPD}^{\bar{m}}(t,f) = \angle Y_{m_1}(t,f) - \angle Y_{m_2}(t,f), \tag{1}$$

where $\angle Y_i$ denotes the phase of the complex spectrogram obtained from the i-th channel. To indicate the look direction θ of the microphone array, the direction feature (DF) [16] is computed as follows:

$$d_\theta(t,f) = \sum_{\bar{m}=1}^{M} \cos(\angle v_\theta^{\bar{m}}(f) - \text{IPD}^{\bar{m}}(t,f)), \tag{2}$$

where $\angle v_\theta^{\bar{m}}(f) = 2\pi f \Delta^{\bar{m}} \cos(\theta^{\bar{m}})/c$ denotes the phase differences between the selected microphone pair for direction θ at frequency f, $\Delta^{\bar{m}}$ is the distance between the selected microphone pair, $\theta^{\bar{m}}$ is the relative angle between the look direction and the microphone pair, and c is the sound velocity. By comparing the phase difference between the steering vector and the IPDs, DF can be used to indicate the sound intensity from the look direction – that is, if the dominant source is from direction θ, then the components of d_θ are close to 1. Finally, the LPS, IPD and DF are concatenated as the input to the separation blocks, which consist of stacked convolution layers with exponential growing dilation factors [17, 18]. At the output, the masks are estimated to extract the complex spectrograms, and an inverse STFT convolution layer is used subsequently to obtain the separated waveforms.

2.3. Multi-look direction based speech separator

In the above method, the look direction needs to be used as the prior information. However, in noisy environments, estimating the accurate direction of arrivals (DOAs) of the speakers becomes challenging. To tackle this problem, a multi-look direction based method proposed in our previous work [6, 8] can be applied for speech separation without the prior DOA information . As its name suggests, several look directions $\{\Theta\}$ are selected to cover the panorama from 0 to 2π, where the network can be regarded as a set of beamformers with different main lobe directions. In the following experiment, we select 4 look directions in the horizontal plane to compute the DF, that is, $\{d(\theta)|\forall\theta \in \{0, 0.5\pi, \pi, 1.5\pi\}\}$.

Two separation networks are developed with the multi-look features. The first separator has 2 outputs corresponding to the signals from the target speakers and the interference speakers respectively. If the input mixture has no interference speaker, the separator only outputs the signals from target speaker. During training, permutation invariant training (PIT) is applied to compute the loss of speech separation, that is,

$$loss_{ss} = \tfrac{1}{2}\min(L(\hat{x}_1, x_t) + L(\hat{x}_2, x_i), L(\hat{x}_2, x_t) + L(\hat{x}_1, x_i)), \tag{3}$$

where x_t, x_i is the signals from the target speaker and interference speaker, $\hat{x}_i$ is the estimated signals from the i-th output, and $L(\cdot)$ is the loss function to compute the difference between the signals. We call this system as Multi-look PIT (MLPIT) separator.

The second separator has the same number of the outputs as that of the look directions, where each output is expected to focus on the closest speaker according to its look direction. The training loss can be computed as:

$$loss_{ss} = \frac{1}{4}\sum_{k=1}^{4} L(\hat{x}_k, x_{\bar{k}}), \tag{4}$$

where the target signals $x_{\bar{k}}$ are selected from $\{x_t, x_i\}$ based on the look direction Θ_k. We call this system as ML enhancement network (MLENet).

In Eq.(3) and Eq.(4), the scale-invariant signal-to-noise(SI-SNR) is used as the loss function, which is defined as:

$$L(\hat{x}, x) = 10\log_{10}\frac{\|x_{target}\|_2^2}{\|e_{noise}\|_2^2}, \tag{5}$$

where $x_{target} = \frac{<\hat{x}, x>x}{\|x\|_2^2}$ and $e_{noise} = \hat{x} - x_{target}$.

2.4. Joint training of speech separator and embedding extractor

To tightly integrate the speech separators and speaker embedding extractors, a multi-task objective loss function is employed, which can be expressed as:

$$loss = \alpha * loss_{ss} + loss_{sv}, \tag{6}$$

where α is the weight of the speech separation loss $loss_{ss}$, and $loss_{sv}$ is the classification loss of SV. For measuring the similarity between the enrollment speech and test speech, we select the maximum of the cosine similarity scores (CS) of each embedding pair as the final score, that is,

$$score(x_{test}, x_{enroll}) = max\{\text{CS}(emb_i^{test}, emb_j^{enroll})|\forall i, j\}, \quad (7)$$

where emb_i is the embedding extracted from the i-th ouput waveform.

3. EXPERIMENTAL SETUP

Individual speaker embedding extractor training: The individual speaker embedding extractor in SV network is trained on Voxceleb2 development dataset [19] with 5994 speakers, where the utterances are augmented with the noise from MUSAN [20] and reverberation from simulated room impulse response (RIR). The noise and reverberation are added online to the original data, where the SNR range for generic noise and music are set to 5-15dB, and for babble noise set to 13-20dB. The margin in AAM-softmax loss function is set to 0.3. The inputs features are 40-dimensional log-Mel filterbanks, and the instance normalization [21] is applied at the input layer. The Voxceleb1 [22] test dataset is used for evaluation. With the Adam optimizer (learning rate is set to 1e-3), the EER can reach 1.62%.

Individual speech separator training: Since the utterances in Voxceleb datasets are recorded in the wild, the simulated datasets on Voxceleb may not be suitable to compute the SI-SNR for the speech separation. The LibriSpeech training dataset [23] is used to simulate the overlapped speech to train the speech separators, and its development dataset is used to evaluate the SI-SNR performances. Each utterance in the multi-channel datasets is mixed among up to two speakers' utterances and one kind of environment noise, where the reverberation impulse responses with different DOAs are applied to these components. The room reverberations are simulated based on the image method [24] on a 6-element uniform circular array of radius 0.035m with different room configurations.[1] 257-dimensional LPSs are extracted from the spectrograms with 512-length window and 50% hop ratio. IPDs are computed from the 6 microphone pairs, i.e., (1, 4), (2, 5), (3, 6), (1, 2), (3, 4) and (5, 6). The networks are trained on short segments with fixed 3-second length. The ratio of single-speaker utterances to two-speaker utterances is around 1:4. In two-speaker cases, the signal-to-interference ratio (SIR) ranges from -6 dB to 6dB. The signal-to-noise ratio (SNR) ranges from 12dB to 30dB across all the cases.

Joint training: The training dataset for joint training is simulated from Voxceleb2 development set. For each speaker, we select 30 utterances as the subset to simulate the mixtures, and the total number of mixtures for training is around 178K. The test dataset for multi-channel evaluation is simulated from the Voxceleb1 dataset, where the segments of the interference speakers are selected from the Voxceleb1 development dataset, and the segments of the target speakers are selected from the Voxceleb1 test dataset. Thus, the sets of the target speakers and the interference speakers are completely disjoint. We adopt three strategies to perform joint training based on the pre-trained models:

1. Re-train individual speech separator: freeze the embedding extractors and update the parameters of the speech separators.
2. Re-train individual embedding extractor: freeze the speech separators and update the parameters of the embedding extractors.
3. Joint training: update the parameters of the whole framework.

[1]For more details, please refer to [8].

4. RESULTS AND ANALYSIS

4.1. Speech separation performance

In Table 1, we show the speech separation results of the individually trained speech separators. Besides MLPIT and MLENet, we also develop an upper bound separator using the oracle source DOAs to compute the input spatial features DF, which is named direction-aware network (DANet). The average SI-SNRs are computed as the average performance from all output channels. Since the separators can have multiple outputs, for fairness, we also compute the best SI-SNRs by selecting the best one from the channels. The results show that DANet with oracle DOAs can obtain the best performance for speech separation, and MLPIT and MLNet can obtain the similar performance .

Table 1. SI-SNR(dB) evaluation based on simulated the LibriSpeech development dataset

Dataset	Method	Avg.	Best
-	Raw mixtures	2.20	-
LibriSpeech	MLPIT	11.48	13.14
	MLENet	11.78	13.53
	DANet with oracle source DOAs	12.36	14.08

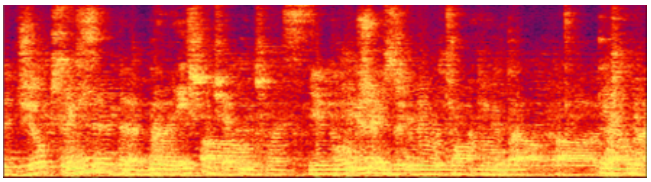
4.2. Performance of the proposed joint training framework

The SV performance is evaluated in terms of EER and minimum Detection Cost Function (mDCF) [25] with $P_{target} = 0.01$. From the results shown in Table 2, we first observe that the pipeline systems without any re-training give the worst performance, which is even worse than an individual SV system without any separator front-end. However, after joint training, the integrated systems can achieve significant performance improvement. Among the three joint training strategies, we find that re-training speaker embedding extractors can only give limited improvement compared with other two strategies. Across all the separators, the best performance in EER is obtained by joint training of the whole framework, where the relative EER reductions of the joint training systems can reach around 52% and 61% compared to the individual SV system and pipeline systems respectively. It is also interesting to find that only re-training the separators can obtain the best performance in mDCF for MLPIT and DANet, which implies that re-training the front-end network is more effective than re-training the back-end networks in overlapped SV tasks.

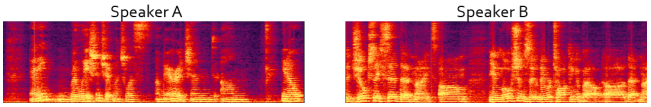
To attempt a deep investigation into joint training, Figure 3 shows an example of the separated speech spectrograms before and after joint training, from which we can draw some inferences about the worse performance of the pipeline systems. Figure 3(a) shows the spectrograms of the overlapped speech from two speakers, and Figure 3(b) shows the corresponding clean spectrogram for each speaker respectively. The spectrograms of the separated speech extracted from the mixture are shown in Figure 3(c) with the pre-trained trained MLENet . Compared with Figure 3(b), some low-SIR spectrogram bins are totally removed with leaving black-hole like regions. Considering the pooling layer in the speech embedding extractor, the mean and standard deviation computed from these regions will diverge greatly from the normal distribution, which may lead to significant degradation in SV performance. However, after joint training, the holes are filled with noise-like spectrogram with small energy in Figure 3(d), and the SV performance becomes much better, which implies that the SV network is sensitive to the whole spectrogram even though some regions do not contain the target speaker information.

Table 2. The performance with different joint training strategies: the pipeline systems are the concatenations of the individual separators and embedding extractors (for individual SV system, there is no separator front-end); and re-train separator/extractor only updates the parameters of the separator/embedding extractor networks.

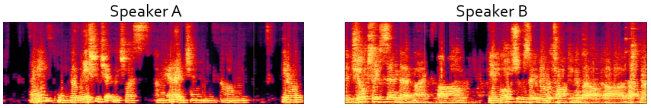
Speech separator	Pipeline		Re-train separator		Re-train extractor		Joint train $\alpha = 0$	
	EER%	mDCF	EER%	mDCF	EER%	mDCF	EER%	mDCF
No separator (individual SV)	17.85	0.701	-	-	-	-	-	-
MLPIT	22.11	0.976	8.76	**0.532**	10.15	0.605	**8.52**	0.542
MLENet	22.85	0.985	9.56	0.571	9.96	0.575	**8.35**	**0.546**
DANet with oracle target DOAs	19.73	0.955	7.65	**0.484**	9.26	0.593	**7.28**	0.520



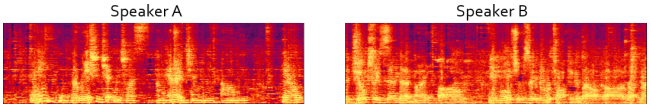
(a) The overlapped speech mixed from two speakers.



(b) The clean speech of two speakers.



(c) The separated speech extracted by the individually trained MLENet.



(d) The separated speech extracted by the MLENet after joint training.

Fig. 3. The spectrograms of the overlapped speech and separated speech from two speakers.

Table 3 shows the performance of the framework with varying parameter α in Eq.(6), where the best mDCF performance of the all three systems can be obtained with $\alpha = 0.5$.

Table 3. System performance with different loss functions.

Speech separator	$\alpha = 0$		$\alpha = 0.5$		$\alpha = 1$	
	EER%	mDCF	EER%	mDCF	EER%	mDCF
MLPIT	8.52	0.542	**8.17**	**0.521**	8.36	0.531
MLENet	**8.35**	**0.546**	8.68	**0.546**	9.07	0.581
DANet	**7.30**	0.520	7.34	**0.497**	7.43	0.504

We also set several evaluation cases considering different DOA distributions, where the angle differences between the target speakers and the interference speakers are limited in fixed ranges. The results are shown in Table 4. When the interference speaker move closer to the target speaker, the performance becomes worse due to the poor separation quality. In the severest case, where the angle differences are smaller than 30°and the spatial information cannot be fully exploited in 6-element microphone array, DANet with oracle target DOAs cannot even beat the other two systems. When the angle differences become larger than 30°, the DANet system returns to the best.

The results in different SIR cases are shown in Table 4, where infinite SIR denotes no interference speaker. In this case, the individual SV system can obtain the best EER due to the noisy data augmentation. When one or more speakers exist, the performance gets better as SIR becomes larger.

Table 4. Results with various angle differences in EER(%) ($\alpha = 0$).

Speech separator	Angle difference			Avg.
	<30°	30°-90°	90°-180°	
No separator	17.74	17.49	17.43	17.56
MLPIT	9.58	7.54	7.73	8.28
MLENet	10.35	7.35	6.97	8.22
DANet	10.58	5.59	5.29	7.15

Table 5. Results with different SIRs in EER(%) ($\alpha = 0$).

Speech separator	-6 dB	0 dB	6 dB	Inf.
No separator	27.83	15.77	7.13	3.06
MLPIT	11.80	7.62	5.79	3.32
MLENet	11.38	7.58	5.71	3.71
DANet	9.60	6.76	4.95	3.45

4.3. Robustness of the proposed joint training framework

In some applications, the enrollment speech may be stored in a remote database and not in multi-channel format. In this section, we construct another case to evaluate the robustness of the re-trained embedding extractors in joint training framework ($\alpha = 0$), where the enrollment recordings are single-channel, while the test recordings are multi-channel. The results are shown in Table 6. Compared with the last columns in Table 5 and Table 2, the EERs in one-speaker and two-speaker cases are very close, which means that the speaker embedding extractors after joint training can still work in the original single-channel cases.

Table 6. Performance evaluation when enrollment speech is single-channel and test speech is multi-channel in EER(%).

Speech separator	One speaker	Two speakers
MLPIT	3.50	8.86
MLENet	3.67	8.44
DANet	3.72	7.53

5. CONCLUSION

In this paper, we present a joint training framework of multi-look separator and speaker extractor for overlapped speech in the speaker verification task. Several joint training strategies are investigated. The experimental results show that joint training can significantly improve the performance of the individual SV system by around 52% relative EER reduction and 23% mDCF reduction, and the re-training of the separator is more effective than the re-training of the embedding extractor.

6. ACKNOWLEDGEMENTS

This project is partially supported by the HKSARG Research Grants Council's Theme-based Research Scheme (Project No. T45-407/19N).

7. REFERENCES

[1] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[2] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[3] Xuankai Chang, Wangyou Zhang, Yanmin Qian, Jonathan Le Roux, and Shinji Watanabe, "Mimo-speech: End-to-end multi-channel multi-speaker speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 237–244.

[4] Tara N Sainath, Ron J Weiss, Kevin W Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michiel Bacchiani, Izhak Shafran, Andrew Senior, Kean Chin, et al., "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.

[5] Mehrez Souden, Shoko Araki, Keisuke Kinoshita, Tomohiro Nakatani, and Hiroshi Sawada, "A multichannel mmse-based framework for speech source separation and noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2013.

[6] Xuan Ji, Meng Yu, Jie Chen, Jimeng Zheng, Dan Su, and Dong Yu, "Integration of multi-look beamformers for multi-channel keyword spotting," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7464–7468.

[7] Jianwei Yu, Bo Wu, Rongzhi Gu Shi-Xiong Zhang Lianwu Chen Yu, Yong Xu Meng, Dan Su, Dong Yu, Xunying Liu, and Helen Meng, "Audio-visual multi-channel recognition of overlapped speech," in *INTERSPEECH*, 2020.

[8] Meng Yu, Xuan Ji, Bo Wu, Dan Su, and Dong Yu, "End-to-end multi-look keyword spotting," in *INTERSPEECH*, 2020.

[9] Ladislav Mošner, Pavel Matějka, Ondřej Novotnỳ, and Jan Honza Černockỳ, "Dereverberation and beamforming in far-field speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5254–5258.

[10] Ladislav Mošner, Oldřich Plchot, Johan Rohdin, Lukáš Burget, and Jan Černockỳ, "Speaker verification with application-aware beamforming," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 411–418.

[11] Wei Rao, Chenglin Xu, Eng Siong Chng, and Haizhou Li, "Target speaker extraction for multi-talker speaker verification," in *INTERSPEECH*, 2019.

[12] Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Leibny Paola García-Perera, Fred Richardson, Réda Dehak, et al., "State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations," *Computer Speech & Language*, vol. 60, pp. 101026, 2020.

[13] Aleksei Gusev, Vladimir Volokhov, Tseren Andzhukaev, Sergey Novoselov, Galina Lavrentyeva, Marina Volkova, Alice Gazizullina, Andrey Shulipa, Artem Gorlanov, Anastasia Avdeeva, et al., "Deep speaker embeddings for far-field speaker recognition on short utterances," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 179–186.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[16] Zhuo Chen, Xiong Xiao, Takuya Yoshioka, Hakan Erdogan, Jinyu Li, and Yifan Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 558–565.

[17] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[18] Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Lianwu Chen, Yuexian Zou, and Dong Yu, "Multi-modal multi-channel target speech separation," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[19] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[20] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[21] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2017.

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[23] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[24] Jont B Allen and David 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.

[25] Seyed Omid Sadjadi, Timothée Kheyrkhah, Audrey Tong, Craig S Greenberg, Douglas A Reynolds, Elliot Singer, Lisa P Mason, and Jaime Hernandez-Cordero, "The 2016 nist speaker recognition evaluation.," in *Interspeech*, 2017, pp. 1353–1357.