INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Multi-task Neural Network for Robust Multiple Speaker Embedding Extraction

Weipeng He[1], *Petr Motlicek*[1], *Jean-Marc Odobez*[1,2]

[1]Idiap Research Institute, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{weipeng.he, petr.motlicek, odobez}@idiap.ch

Abstract

This paper introduces a novel approach for extracting speaker embeddings from audio mixtures of multiple overlapping voices. This approach is based on a multi-task neural network. The network first extracts a latent feature for each direction. This feature is used for detecting sound sources as well as identifying speakers. In contrast to traditional approaches, the proposed method does not rely on explicit sound source separation. The neural network model learns from data to extract the most suitable features of the sounds at different directions. The experiments using audio recordings of overlapping sound sources show that the proposed approach outperforms a beamforming-based traditional method.

Index Terms: Multi-task learning, speaker embedding, speaker verification, microphone array processing

1. Introduction

Speaker recognition is an important technology for many applications, such as robots and smart speakers. Knowing speaker identities allows natural long-term interactions, as well as speaker-adapted automatic speech recognition. In real situations, robots and smart speakers may operate in noisy and dynamic environments with overlapping sounds. In this paper, we investigate speaker recognition under such a condition.

Speaker recognition includes two different tasks: *speaker identification* and *speaker verification* [1]. Speaker identification aims to identify an unknown speaker from a set of known speakers, whereas speaker verification aims to verify if a voice and some enrolled voices are from the same speaker. Although the goals of these tasks are different, the techniques they rely on are similar. In fact, most of the speaker recognition methods are based on mapping speech segments to a speaker embedding space where they can be compared using a metric for identification or verification. In an ideal embedding space, distances between voices of the same speaker are smaller than distances between voices of different speakers.

Speaker recognition with clean and segmented single-channel audio has been extensively studied. Well-known approaches include *Gaussian Mixture Model* (GMM) [2], GMM with *Universal Background Model* (UBM) [3], *Joint Factor Analysis* (JFA) [4], *Support Vector Machine* (SVM) for GMM supervector classification [5], and the *i-vector* system [6]. Recently, many deep learning based approaches have been shown to outperform the traditional ones [7–14]. These deep learning based approaches extract speaker embeddings in two ways. One way is to train a network for speaker identification and use the activation at one of the last hidden layers as speaker embeddings [7, 13]. In contrast, the other way is to use directly the network output as the speaker embedding, and train the network with objective functions that are defined on the distances between same-speaker and different-speaker pairs. Examples of the objective functions include *contrastive loss* [8], which separately minimizes distances between same-speaker pairs and maximizes those between different-speaker pairs, and *triplet loss* [9, 14], which maximizes the difference between different-speaker distances and same-speaker distances up to a given margin.

Besides speaker recognition from clean audio signals, a number of studies address speaker recognition in the presence of noise and simultaneous speakers. These approaches rely on separating the sound sources (from either single-channel or multi-channel audio signals), so that speaker recognition is applied on separated single-channel signals. Sound separation is applied prior and independently to the speaker recognition in the *sequential approaches* [15–18]. Alternatively, sound separation and speaker recognition are solved jointly in the *joint approaches* [19–21]. Nevertheless, using deep neural networks for speaker recognition under the multi-speaker condition is still an emerging topic. Specifically, joint *Direction-of-Arrival* (DOA) estimation and recognition of multiple speakers has not been studied so far.

This paper investigates deep neural networks for speaker recognition under the multi-speaker condition using DOA estimation as an auxiliary task. We use the neural networks to extract features for each direction, which are shared for both DOA estimation and speaker embedding. In contrast to previous works, our approach does not rely on explicit separation of the signals. Instead, the network learns to implicitly separate the sound features through end-to-end training.

Our proposed neural network shares similarities with the well-known X-vector network [13], that both networks are trained using speaker identification loss and extract speaker embeddings from hidden layers. Moreover, temporal statistic pooling is used in both approaches to accommodate input sequences of variable lengths. The difference between our approach and the X-vector approach is that we address speaker recognition of multiple overlapping speakers from multi-channel audio, while X-vector is designed for single-channel single-speaker audio.

2. Approach

We describe our multi-task neural network approach in terms of input representation, network output, loss function, and network architecture.

2.1. Network Input

We use the raw *Short-Time Fourier Transform* (STFT) as the network input. STFT includes both the spectral power information as well as the phase information of the input signal.

http://dx.doi.org/10.21437/Interspeech.2021-1769

For DOA estimation, *Inter-channel Level Difference* (ILD) and spectral cues can be extracted from the power information, and *Inter-channel Phase Difference* (IPD) can be extracted from the phase information. The power information, in addition, includes necessary features for speaker recognition.

The STFT is processed in the same way as in [22–24], except that the input segment can be arbitrarily long to incorporate more information for speaker recognition. Specifically, the input audio signals captured by a microphone array contain four audio channels and are sampled at 48 kHz. STFT is extracted from the input audio signals using frames of 2048 samples (43 ms) with 50% overlap. The 337 frequency bins between 100 and 8000 Hz are used. The real and imaginary parts of the STFT coefficients are split into two individual channels. Therefore, the input feature of each unit has a dimension of $T \times 337 \times 8$, where the number of frames T varies across different segments.

2.2. Network Output and Loss Function

The network output includes frame-wise prediction of the spatial spectrum $\mathbf{p}_t = \{p_{td}\}_{d=1}^{D} \in [0,1]^D$ for DOA estimation, and segment-wise prediction of speaker posterior probability at each direction $\mathbf{q}_d = \{q_{ds}\}_{s=1}^{S} \in [0,1]^D$ for speaker identification. The subscripts $t \in \{1, 2, \ldots, T_o\}$ is the frame index, $d \in \{1, 2, \ldots, D\}$ is the direction index, and $s \in \{1, 2, \ldots, S\}$ is the speaker ID. Due to downsampling, the frame rate of predicted spatial spectrum is different from that of the input, thus $T_o \neq T$.

Encoding. The desired output spatial spectrum is encoded by the Gaussian-based spatial spectrum coding [25], that is:

$$p_{td} = \begin{cases} \max_{\varphi \in y_t} \left\{ e^{-\delta(\varphi_d, \varphi)^2 / \sigma^2} \right\} & \text{if } |y_t| > 0 \\ 0 & \text{otherwise} \end{cases}, \quad (1)$$

where $y_t \subset \Phi$ is the set of ground truth directions at frame t, σ is the parameter to control the width of the Gaussian curves, $\delta(\cdot, \cdot)$ denotes the azimuth angular distance, and $|\cdot|$ denotes the cardinality of a set.

Inspired by how sound type is encoded in [22], the speaker ID prediction at direction φ_d depends on the nearest sound source (speaker) to that direction, that is:

$$q_{ds} = \begin{cases} 1 & \text{if Speaker } s \text{ is the nearest speaker to } \varphi_d \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

Loss Functions. The target loss function is a linear combination of the individual task-specific loss functions:

$$\text{Loss} = \mu \, \text{Loss}_{DOA} + \lambda \, \text{Loss}_{ID}, \quad (3)$$

where μ and λ are weighting parameters. We use the *Mean Squared Error* (MSE) loss for DOA estimation:

$$\text{Loss}_{DOA} = \frac{1}{T_o} \sum_{t=1}^{T_o} \|\hat{\mathbf{p}}_t - \mathbf{p}_t\|_2^2, \quad (4)$$

where $\hat{\mathbf{p}}_t$ and $\mathbf{p}_t$ are the actual and desired spatial spectrum outputs, respectively. The speaker identification loss is the weighted sum of cross entropy loss at individual directions:

$$\text{Loss}_{ID} = -\sum_{d=1}^{D} w_d \sum_{s=1}^{S} q_{ds} \log \hat{q}_{ds}, \quad (5)$$

where $\hat{q}_{ds}$ and q_{ds} are the actual and desired speaker identity outputs, respectively. The weighting $\{w_d\}$ depends on its distance to the DOAs of the sound sources:

$$w_d = \begin{cases} \max_{\varphi \in y} \left\{ e^{-\delta(\varphi_d, \varphi)^2 / \sigma_w^2} \right\} & \text{if } |y| > 0 \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

where $y = \cup_t y_t$ contains the segment-level ground truth directions.

Decoding. During test time, the network outputs frame-wise spatial spectra $\mathbf{p}_t$ and speaker embedding $\mathbf{r}_d$ per direction (will be explained in Section 2.3). To get segment-level DOA prediction, we compute the average of frame-wise spatial spectra:

$$\mathbf{p} = \frac{1}{T_o} \sum_{t=1}^{T_o} \mathbf{p}_t \quad (7)$$

and apply peak finding according to detect sound sources. For any detected sound source, the speaker embedding output at the estimated direction is the predicted speaker embedding.

2.3. Network Architecture

We design a multi-task network for speaker embedding using DOA estimation as an auxiliary task. Its architecture, depicted in Fig. 1, consists of a trunk for feature extraction, and two task-specific branches. The trunk (green blocks) applies 2D convolutions along time and frequency axes to extract *Time-Frequency* (TF) local features. It starts with two downsampling convolutions to reduce the computational cost. They are followed by five residual blocks, which are used for extracting high-level TF-local features, each of which is a 480-dimensional vector. Each of these feature is then separated into DOA-wise features at $D = 120$ directions (4-dimensional vector per direction). Then, they are re-organized by merging features across all frequencies (54 bins after down-sampling). As a result, the trunk extracts time-DOA local features, each of which is a 216-dimensional vector ($216 = 4 \times 54$). These features are then used as input for the task-specific branches.

The DOA estimation branch (blue blocks) applies two layers of 2D convolutions along time and DOA axes. The borders are padded circularly along the DOA axis, preserving its actual topology. This branch outputs one value per direction per frame, which is bounded between 0 and 1 by the sigmoid function. This output is the frame-wise spatial spectrum $\mathbf{p}_t$.

The speaker recognition branch (red blocks) starts with two layers of 2D convolutions to extract frame-wise speaker features per direction $\mathbf{f}_{td} \in \mathbb{R}^{512}$, which is then pooled along the time axis using *weighted average and standard deviation*:

$$\mathbf{f}_d^{(avg)} = \frac{\sum_{t=1}^{T_o} p_{td} \mathbf{f}_{td}}{\sum_{t=1}^{T_o} p_{td}}, \quad (8)$$

$$\mathbf{f}_d^{(std)} = \sqrt{\frac{\sum_{t=1}^{T_o} p_{td} \left(\mathbf{f}_{td} - \mathbf{f}_d^{(avg)} \right)^2}{\sum_{t=1}^{T_o} p_{td}}}, \quad (9)$$

where $\sqrt{\cdot}$ and $\cdot^2$ are element-wise square root and square, respectively. As indicated in the formulas, we use the output of the DOA estimation branch $\{p_{td}\}$ as the weighting parameters, because the DOA estimation output (i.e. spatial spectrum) indicates whether there is an active sound at that frame and direction. This can be viewed as an attention mechanism, that is the network chooses by itself which frames to attend to.

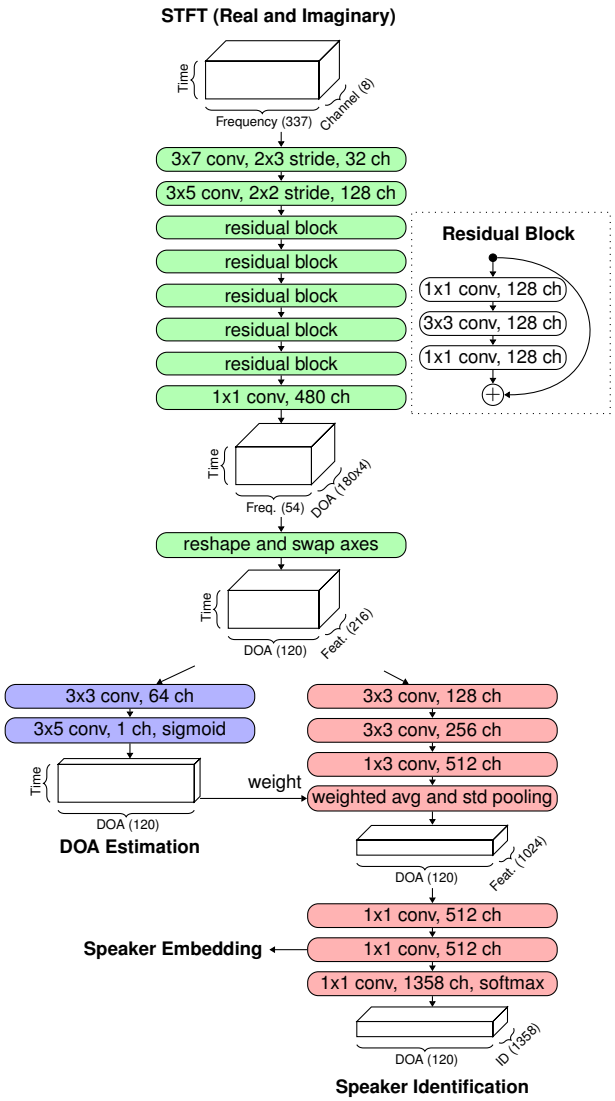

Figure 1: *The architecture of the multi-task network for speaker recognition.*

Their concatenation $\mathbf{f}_d = [\mathbf{f}_d^{(avg)} \; \mathbf{f}_d^{(std)}] \in \mathbb{R}^{1024}$ is the segment-level speaker feature per direction. Then, the speaker identity posterior probability is computed from these features with fully-connected layers (1×1 convolutions) and a softmax layer.

At test time, the 512-dimensional activation of the last hidden layer after batch normalization is the speaker embedding $\mathbf{r}_d$ at the direction φ_d.

3. Experiments

We compared the proposed approach to a sequential approach using real robot audio data in a speaker verification setting.

3.1. Data

We used the loudspeaker recordings from SSLR dataset [1] [25] for training and evaluation. The data includes audio recordings of speech signals played from loudspeakers and recorded by a Softbank Pepper robot. The robot has four co-planar microphones on its head. There are up to two overlapping sound sources. The speech data were selected from the AMI corpus [26]. The specifications of the loudspeaker data are listed in Table 1.

[1] https://www.idiap.ch/dataset/sslr

Table 1: *Specifications of the recorded data.*

	Training	Test
# of files	4208	2393
- single source	2808	1597
- two sources	1400	796
# of male speakers	105	8
# of female speakers	43	8
Total duration	16 hours	8 hours
Azimuth (°)	$[-180, 180]$	$[-180, 180]$
Elevation (°)	$[-39, 56]$	$[-29, 45]$
Distance (m)	$[0.5, 1.8]$	$[0.5, 1.9]$

Training models that identify speakers per direction requires more variability in DOAs of individual speaker as well and number of identities. Otherwise, the network may overfit to a wrong state where spatial locations are used as the clues for speakers' identity. Therefore, we added simulated data to complement the real loudspeaker recordings for training. We used the RIR-Generator [27] to generate room impulse responses. The impulse responses are convolved with clean source signals randomly selected from the VoxCeleb1 dataset [28]. The spatialized audio samples are mixed to generate (up to four) overlapping sounds and added with real robot background noise (fan noise). In total, there are 1358 speakers (147 from the loudspeaker data and 1211 for simulated data) for training and 16 different speakers (8 male and 8 female) for evaluation.

3.2. Training Process and Parameters

The training processing includes two steps. First, the model is trained with an emphasis on DOA estimation ($\mu = 1, \lambda = 0.1$) and 3-10 second training segments. Then, the second step, it is trained with an emphasis on speaker identification ($\mu = 0.1$, $\lambda = 1$) and 2-5 second training segments. In each mini-batch, 10.0% of the sequences are sampled from the loudspeaker dataset, and the rest are sampled from the simulated dataset. We select the number of sequences in each mini-batch, such that they approximately fill up the memory of a GPU with 11 GB memory. Thus, depending on the sequence length, the number of sequences in each mini-batch varies between 10 and 90. The model is trained for 80 epochs with an Adam optimizer [29] at each step. The learning rate is 0.001 for the first 40 epochs and reduced by half for the other 40 epochs. Other parameters are chosen as $\sigma = 8°$ and $\sigma_w = 16°$ in the experiments. This proposed method is denoted by PROP.

3.3. Baseline Sequential Method

We use a sequential method (denoted by SEQ) as the baseline for comparison. This method is based on the *Minimum Variance Distortionless Response* (MVDR) beamformer [30] and a deep neural network for speaker embedding [14]. The neural network directly output speaker embedding using single-channel audio input. It is first trained on the VoxCeleb1 dataset [28] with a triplet loss and intra-class distance variance regularization, and then fine-tuned using the beamformed signals extracted from the loudspeaker training data according to the ground truth DOAs.

Table 2: *EER (%) on trials with pairs of sound sources from all segments. The second column indicates how the DOA used for speaker embedding extraction is obtained.*

Method	DOA	Segment duration			
		10s	5s	3s	2s
SEQ	GT	8.72	11.07	13.48	18.90
-	EST	8.67	11.04	**13.49**	18.93
PROP	GT	5.91	9.76	13.42	17.80
-	EST	**5.97**	**9.83**	13.50	**17.85**
-	EMT	6.96	10.85	14.32	18.63

Table 3: *EER (%) under different trial pair conditions. The speaker embedding are extracted at directions estimated by the single-task network (EST).*

Trial	Method	Segment duration			
		10s	5s	3s	2s
1 vs 1	SEQ	**3.80**	**7.07**	**10.19**	**16.42**
-	PROP	5.24	8.83	12.50	16.88
1 vs m	SEQ	10.82	12.73	15.23	20.40
-	PROP	**6.36**	**10.22**	**13.99**	**18.43**
m vs m	SEQ	12.24	14.29	16.87	21.70
-	PROP	**6.34**	**10.76**	**14.71**	**19.18**

3.4. Speaker Verification Performance

We evaluate the *Equal Error Rate* (EER) of the speaker embedding methods under a speaker verification setting (Table 2). First the loudspeaker test data are segmented into 2, 3, 5, or 10-second segments. For each sound source in each segment, we extract the speaker embedding under different DOA conditions:

- **GT**: ground truth DOA is known;
- **EMT**: DOA is estimated by the proposed multi-task network;
- **EST**: DOA is estimated by a single-task network from [24].

Then, we generate verification trials using up to 5 million randomly-sampled sound source pairs. We compute the cosine similarity scores between the speaker embeddings of all trial pairs. Comparing the scores to a threshold, we can make predictions on whether the speakers from a trial pair have the same identity. The EER is the rate when false acceptance rate and false rejection rate are equal while varying the threshold.

In addition to segment duration, speaker verification are also strongly affected by whether there are overlapping sounds or not. Therefore, we additionally report the speaker verification performance on trail pairs of these different conditions (Table 3):

- **1 vs 1**: Both speaker embeddings are sampled from the single-source segments;
- **1 vs m**: One is from the single-source segments and the other is from the multi-source segments;
- **m vs m**: Both are from the multi-source segments.

In this table, both the proposed and the baseline systems use DOA estimation based on the single-task network (EST), as this is the best condition that can be achieved in real applications. Results under other DOA conditions (GT and EMT) are omitted for brevity.

DOA estimation. Comparing the results of a method with different DOA estimation, the speaker embeddings extracted from the directions predicted by the single-task DOA estimation model (EST) is as good as using the ground truth (GT). This indicates that the prediction of the DOA estimation model is accurate and the speaker embedding approaches are robust to small error in DOA estimation. However, the performance is degraded when an inaccurate DOA estimation is used for speaker embedding extraction. This is the case if the DOA estimation of multi-task model (EMT) is used. This is expected as we consider the DOA estimation as an auxiliary tasks. The parameters of DOA estimation branch are tuned to produce the best temporal weighting for speaker recognition instead of to get best DOA estimation scores.

The results indicate that for practical applications the optimal solution is to run two neural networks parallelly: using the proposed multi-task network for speaker embedding extraction at all directions and the single-task network to estimate which direction should be used to pick the speaker embeddings. Our experiments using a mainstream GPU show that the extra computational cost is negligible.

Proposed vs. sequential methods. The proposed method, compared to the sequential method, achieves better overall performance in long segments (5 and 10-second segments), while their EERs in short segments (2 and 3-second segments) are similar. Their performance under different trial conditions (Table 3) indicates that while the proposed method is not as good as extracting speaker embedding in single-source segments, it is significantly better under the multi-source condition. In the single-source case, the sequential approach is not influenced much by the beamformer, as sound source separation is not necessary and the single-channel speaker embedding network can be trained to handle noisy input (as what the fine-tuning is for). In contrast, our proposed multi-task network aims to extract speaker embeddings on all directions, and is more complex than a single-channel single-speaker embedding network. Therefore, it is more difficult to train. In our experiments, we find that the single-channel speaker embedding approach is more suitable for single-source recordings. However, for segments containing multiple sound sources, the sequential approach relies on the beamformer to separate the signals. Its speaker embedding performance may degrade due to imperfect sound separation, whereas our proposed approach does not require explicit sound separation.

4. Conclusion

In this paper, we have presented a novel multi-task neural network for extracting speaker embeddings of multiple simultaneous speakers using DOA estimation as an auxiliary task. The network learns to estimate a spatial spectrum score and a speaker embedding for each direction. The spatial spectrum is used as weighting parameters for weighted average and standard deviation pooling of the frame-wise speaker features along the time axis. Compared to a sequential approach that applies separately DOA estimation, beamforming and speaker embedding extraction, our proposed approaches achieves better overall performance for audio segments with overlapping sound sources.

5. Acknowledgements

This research was funded by the European Commission Horizon 2020 program under the projects MuMMER (multimodal mall entertainment robot, grant agreement ID 688147) and ROXANNE (real time network, text, and speaker analytics for combating organized crime, grant agreement ID 833635).

6. References

[1] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, Nov. 2015.

[2] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, Jan. 2000.

[4] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," CRIM, Montreal, Quebec, Canada, Tech. Rep. CRIM-06/08-13, 2005.

[5] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

[6] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[7] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4052–4056.

[8] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2016, pp. 165–170.

[9] H. Bredin, "TristouNet: Triplet loss for speaker turn embedding," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 5430–5434.

[10] S. Dey, P. Motlicek, S. Madikeri, and M. Ferras, "Exploiting sequence information for text-dependent speaker verification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5370–5374.

[11] S. Dey, T. Koshinaka, P. Motlicek, and S. Madikeri, "DNN based speaker embedding using content information for text-dependent speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5344–5348.

[12] S. Dey, S. R. Madikeri, and P. Motlicek, "End-to-end Text-dependent Speaker Verification Using Novel Distance Measures." in *INTERSPEECH 2018*, 2018, pp. 3598–3602.

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 5329–5333.

[14] N. Le and J.-M. Odobez, "Robust and discriminative speaker embedding via intra-class distance variance regularization." in *Interspeech 2018*, 2018, pp. 2257–2261.

[15] T. May, S. van de Par, and A. Kohlrausch, "Binaural localization and detection of speakers in complex acoustic scenes," in *The Technology of Binaural Listening*, ser. Modern Acoustics and Signal Processing, J. Blauert, Ed. Berlin, Heidelberg: Springer, 2013, pp. 397–425.

[16] X. Zhao, Y. Shao, and D. Wang, "CASA-based robust speaker identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1608–1616, Jul. 2012.

[17] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 836–845, Apr. 2014.

[18] H. Taherian, Z.-Q. Wang, J. Chang, and D. Wang, "Robust Speaker Recognition Based on Single-Channel and Multi-Channel Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1293–1302, 2020.

[19] J. Zegers and H. Van hamme, "Joint sound source separation and speaker recognition," in *Interspeech 2016*, Sep. 2016, pp. 2228–2232.

[20] L. Drude, T. von Neumann, and R. Haeb-Umbach, "Deep attractor networks for speaker re-identification and blind source separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 11–15.

[21] Y. Shi, Q. Huang, and T. Hain, "Speaker re-identification with speaker dependent speech enhancement," *arXiv:2005.07818 [cs, eess]*, Aug. 2020.

[22] W. He, P. Motlicek, and J.-M. Odobez, "Joint localization and classification of multiple sound sources using a multi-task neural network," in *Interspeech 2018*, Sep. 2018, pp. 312–316.

[23] ——, "Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 770–774.

[24] ——, "Neural Network Adaptation and Data Augmentation for Multi-Speaker Direction-of-Arrival Estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1303–1317, 2021.

[25] ——, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 74–79.

[26] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, and others, "The AMI meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.

[27] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[28] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 2616–2620.

[29] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, San Diego, May 2015.

[30] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.