SPEAKER-AWARE TARGET SPEAKER ENHANCEMENT BY JOINTLY LEARNING WITH SPEAKER EMBEDDING EXTRACTION

Xuan Ji[1]*, Meng Yu*[2]*, Chunlei Zhang*[2]*, Dan Su*[1]*, Tao Yu* [3]*, Xiaoyu Liu* [4]*, Dong Yu*[2]

[1]Tencent AI Lab, Shenzhen, China, [2]Tencent AI Lab, Bellevue, WA, USA
[3]Tencent IEG, Bellevue, WA, USA, [4]Tencent IEG, Shenzhen, China

ABSTRACT

Deep learning based speech separation approaches have received great interest, among which the recent speaker-aware speech enhancement methods are promising for solving difficulties such as arbitrary source permutation and unknown number of sources. In this paper, we propose a novel training framework which jointly learns the speaker-conditioned target speaker extraction model and its associated speaker embedding model. The resulting unified model directly learns the appropriate speaker embedding for improved target speech enhancement. We demonstrate, on our large simulated noisy and far-field evaluation sets of overlapped speech signals, that our proposed approach significantly improves the speech enhancement performance compared to the baseline speaker-aware speech enhancement models.

Index Terms: speaker-aware, target speech enhancement, speaker embedding, joint learning

1. INTRODUCTION

Human auditory system has a mechanism for separating mixed signals. Much research attention has been given to the topic of employing the machine to emulate human auditory perception. The progress made in multi-talker mixed speech separation, often referred to as the cocktail-party problem [1], has been less impressive than that in single-talker speech enhancement. More generally, source separation is to invert the unknown mixing process and estimate the individual source signals in cases where a set of source signals of interest goes through an unspecified mixing process and is recorded at a single microphone or a microphone array. Inspired by the success of deep learning on speech recognition [2], researchers developed deep learning based techniques to address the monaural speech separation problem. The performance improvements are particularly impressive with the very recent techniques such as deep clustering (DPCL) [3], deep attractor network (DANet) [4] and permutation invariant training (PIT) [5], which were proposed to address two major issues in the task of blind speech separation, i.e. unknown number of speech sources and permutation of speaker labels in the training, respectively.

A new category of algorithms have been developed recently to extract the voice of a target speaker of interest from the the mixture with all the other speakers and noises. Target speaker enhancement is closely related to classical speech separation tasks, but is different in a way that a speaker-dependent auxiliary feature differentiates the target speaker from the rest in the mixture. For example, text-dependent enhancement model only preserves the speech signal of certain words [6]. Location based spatial information of the target speaker is employed in the multi-channel setting [7, 8]. Visual information such as target speaker's lip movement in a video assists to achieve impressive results [9, 10]. Pre-enrolled speaker profile utterances have been used as the auxiliary signal to extract a particular speaker, e.g. deep extractor [11] and VoiceFilter [12]. All such systems are more practical in real-world scenarios because they are optimized to predict a single target speaker from the mixture, and thus do not require prior knowledge on the number of speakers and avoid the permutation problem.

With the proliferation of smart homes, mobile and automotive devices, speech-based human-machine interaction becomes prevailing. The user-defined, i.e. enrolled speaker-informed, speech enhancement like VoiceFilter improves a certain user's speech quality not only in ambient noises but also in multi-talker conditions. Existing speaker-informed enhancement systems suffer a major limitation that the pre-computed speaker embedding is obtained by a separated speaker verification model. It has been provided by averaging speaker vectors over the enrolled utterances in [12]. The speaker embedding is able to provide a global bias for the target speaker. Later on, as noted in [13] that the local dynamics and the temporal structure in the enrolled utterances may be helpful for accurate separation, an attention mechanism was introduced to compute the local similarity between the enrolled utterances and the input signal, thus allowing segment-level alignment between input mixture and a proper time-variant speaker embedding. However, neither mean pooling of fixed speaker embeddings [12] or a weighted average of fixed speaker embeddings [13] are directly optimized towards minimizing the enhancement loss. These approaches generally suffer from deficiencies such as being not effective enough in capturing the most proper speaker characteristics for the purpose of target speech enhancement.

In this work, we propose a novel, effective yet simple target speech enhancement model, which jointly optimizes the speech enhancement and auxiliary target speaker feature extraction. A speaker verification network is pre-trained and integrated with the enhancement model for joint learning. The speaker model is thus further adapted by a summation of enhancement loss and speaker verification loss, pursuing a suitable speaker embedding for target speaker extraction. Furthermore, most of previous separation and enhancement systems process the audio stream in time-frequency (TF) domain where the phase of the reconstructed signal is less considered, though it is critical for preserving the speech quality [14, 15]. We incorporate the recent encoder-decoder framework named TasNet [16, 17], and directly conduct the end-to-end target speech enhancement in time domain. The proposed system generalizes the TasNet to work in a target speech enhancement framework and be jointly trained with speaker verification model.

The rest of the paper is organized as follows. In Section 2, we introduce our TasNet based speaker-aware enhancement model and speaker verification model for speaker feature extraction. In Section 3, we present details of the proposed end-to-end model architecture

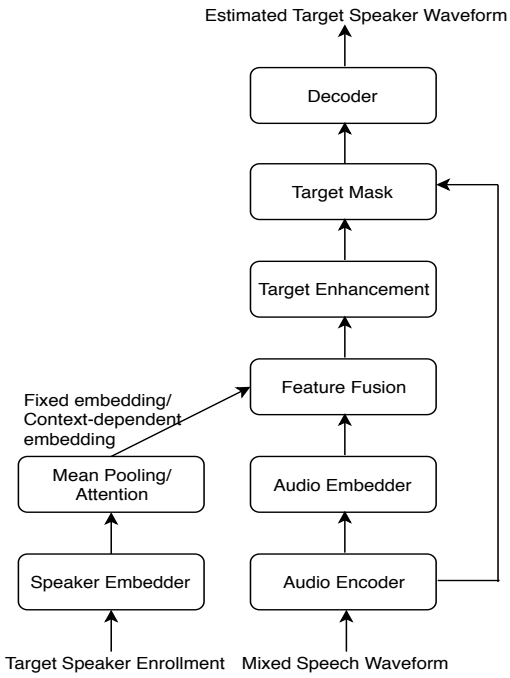


Fig. 1. Speaker-informed time-domain target speech enhancement based on non-unified model structure.

and its joint optimization scheme. We describe our experimental setup in Section 4, and evaluate the effectiveness of the proposed system in Section 5. We conclude this work in Section 6.

2. SPEAKER-INFORMED SPEECH ENHANCEMENT

2.1. Time-domain target speech enhancement model

The task is to extract the target speaker's voice from a noisy and/or mixed speech waveform. We extract the characteristics of the target speaker through its enrollment, which is a set of utterances collected in an enrollment process. The raw waveform of speech mixture and corresponding target speaker embedding are used as the input of the enhancement network which predicts the voice of target speaker. Scale-invariant signal-to-noise ratio (SI-SNR) is used as the training objective function, which is defined as

$$\text{SI-SNR} = 20\log_{10}\frac{\|\alpha\cdot\mathbf{s}_e\|}{\|\mathbf{s}_t-\alpha\cdot\mathbf{s}_e\|}. \quad (1)$$

$\mathbf{s}_e, \mathbf{s}_t$ are estimated signal and target signal, respectively, for which zero-mean normalization is applied. α is an optimal scaling factor computed via

$$\alpha = \mathbf{s}_e^T\mathbf{s}_t/\mathbf{s}_t^T\mathbf{s}_t. \quad (2)$$

The proposed enhancement model is mainly inherited from Conv-TasNet [17], which consists of five parts in our task, an audio encoder, an audio embedder, a feature fusion layer, an enhancement network and a decoder. Given an audio mixture chunk $\mathbf{x_t}$, the audio encoder (enc^a) tends to encode the input samples into a non-negative representation, which goes through an audio embedder (embd^a) and is then concatenated with the speaker embedding ($\text{embd}^v(\mathbf{v}_t)$), where $\mathbf{v}_t$ is the input to the speaker embedder. The enhancement is performed by estimating mask of the target speaker ($\mathbf{m}_t$). The audio

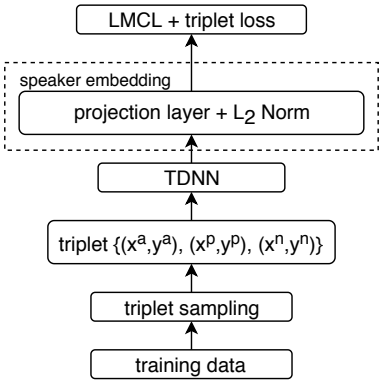


Fig. 2. The speaker embedder model in our system.

decoder is used to reconstruct the masked output into time domain signal. The total framework could be described as

$$\mathbf{s}_t = \text{dec}(\text{enc}^a(\mathbf{x}_t)\odot\mathbf{m}_t). \quad (3)$$

where

$$\mathbf{m}_t = \text{enh}(\text{embd}^a(\text{enc}^a(\mathbf{x}_t)), \text{embd}^v(\mathbf{v}_t)), \quad (4)$$

We keep the implementation of audio encoder/decoder same as the original Conv-TasNet [17], which mainly performs 1D convolution and deconvolution operation (128 40-d kernels) on time domain audio signals and the masked output, respectively. The audio and speaker embeddings are fused in a fusion layer, which is performed through a simple concatenation operation over the channel dimensions, followed by a position-wise projection into a vector of the same dimension (128-d) as the audio embedding. The audio embedder and target enhancement network follow the same implementation as the separation network in [17] except for the single source output mask. Each of the two modules consists of M convolutional blocks with dilation factors $1, 2, 4, \ldots, 2^{M-1}$ repeated R times. We set M to 8 and R to 2 in our experiment.

The TasNet based speaker-informed target speech enhancement model is depicted in Figure 1. Similar to [12], it summarizes the information about the target speaker into a single fixed speaker embedding vector by averaging the speaker embeddings from the enrolled target speaker utterances. No matter what is spoken in the mixed speech, it always uses the same speaker embedding for all frames of the mixed speech. A time-varying context-dependent speaker vector is then proposed in [13] to represent the characteristics of the target speaker. An attention mechanism is applied for dynamically computing the frame-wise speaker feature vector which is a weighted summation of speaker embeddings from the enrolled target speaker. We refer the readers to [13] for the attention implementation in the model. Figure 1 illustrates this context-dependent speaker-informed target speech enhancement model as well by using the attention mechanism in the pooling stage.

2.2. Speaker verification model

Recently, deep neural networks based speaker embedding systems become the state-of-the-art methods for speaker verification (SV) [18, 19, 20, 21]. Speaker embedding outperforms the conventional i-vector in many speaker verification tasks, including text-independent SV such as NIST SREs and Voxceleb [21, 22], as well as text-dependent SV tasks [23]. Inspired by recent studies in face and speaker verification [20, 24, 25, 26], the speaker embedding system

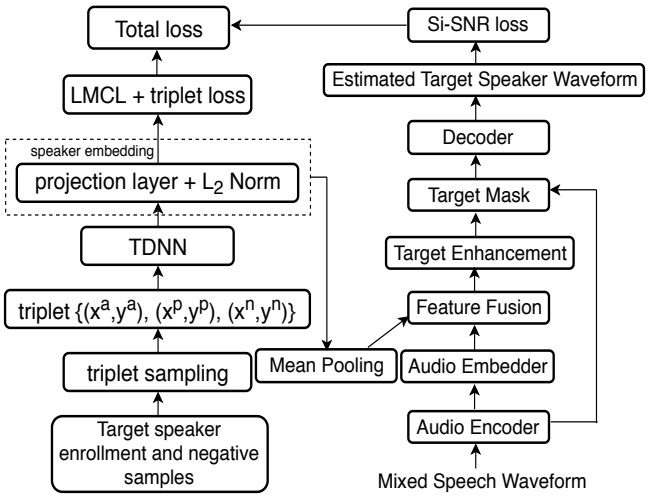


Fig. 3. Joint learning structure of speaker embedder and target speech enhancement.

developed in this study is shown in Figure 2. It corresponds to the "speaker embedder" module in Figure 1.

Unlike Inception Resnet used in [25], we employ a TDNN framework which is similar to [20] for target speech separation task with demanding latency control requirement. Generally, the TDNN utilized to extract speaker embedding consists of frame-level and segment-level layers with a statistical pooling in between. The output of frame-level layers are aggregated across the entire utterance by the pooling layer and further processed by several fully-connected layers. The speaker embedding is usually extracted from a fully-connected layer for downstream tasks. To be more specific, we follow the original TDNN topology with the following modifications. The kernel size is $[5, 5, 7, 1, 1]$. There are 5 TDNN layers in total with 512 channels for the first 4 layers, while the last TDNN layer has the dimension of 1500. After the statistic pooling, the mean and variance vectors are concatenated to form a 3000-d vector, which is followed by a 512-d fully connected layer. Batch-Normalization and ReLU non-linearity are applied to all layers for improved performance and fast convergence. Finally, a linear projection to a 128-d vector is used for speaker embedding extraction, where L_2 norm is applied as an unit energy constraint.

Similar to [25], we use a multi-task objective for speaker embedding training, which consists of a large margin cosine loss (LMCL) [26, 27] and a triplet loss [28] with a tuned weight. The classification losses LMCL always perform well when the training data is large with many speakers while triplet loss improves the performance with relatively small data sets. The total loss is defined below

$$\mathcal{L}_{SV} = \mathcal{L}_{triplet} + \omega_1 \mathcal{L}_{lmc} + \omega_2 \mathcal{L}_r, \tag{5}$$

where $\mathcal{L}_{triplet}$ is triplet loss, $\mathcal{L}_{lmc}$ is LMCL, and $\mathcal{L}_r$ is a L_2 regularization term which alleviates over-fitting during training. Practically, we find $\omega_1 = 0.2$ and $\omega_2 = 0.001$ is a good combination for our experiments.

3. THE JOINT LEARNING FRAMEWORK

The multi-task joint learning has been deployed in speech processing area in recent years which has been known working well to boost correlated tasks. For example, a multi-task LSTM network architecture was proposed in [29], in which a unified objective that considers both the speech enhancement quality and automatic speech recognition (ASR) accuracy is used. A multi-task architecture of speaker verification and speech recognition was proposed in [30] where it focused on speaker adaptation for ASR, while the joint learning framework in [31] employed phonetic information for improved speaker vector alignment in the task of text-dependent speaker verification. The joint training architecture of speaker verification and speech enhancement occurs recently in [32], where a speaker verification loss is used to train the speech enhancement model. By doing this, the enhancement network is expected to select important TF bins for speaker verification, though the quality of speech might not be improved.

Motivated by the multi-task joint learning methods in the past literature, we designed a joint learning architecture for our target speech enhancement as shown in Figure 3. The drawback of the previous architectures of speaker-informed target speech enhancement [12, 13] is that the enhancement step and speaker embedding generation step are independent from each other. As discussed in Section 1, the criterions for the speech enhancement and the speaker verification are different. The speaker embedding extracted from the optimized speaker verification model might not lead to the optimum speech enhancement quality in the enhancement step. The objective function of the proposed network is

$$\mathcal{L}_{Joint} = \mathcal{L}_{SI-SNR} + \alpha \mathcal{L}_{SV}, \tag{6}$$

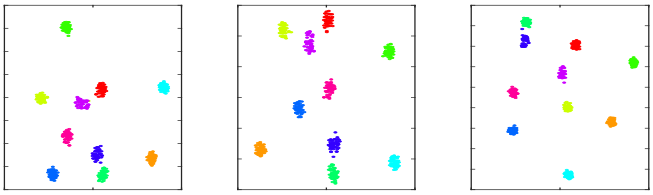
where α, set to 1 in the experiment, is a weight for speaker verification model.

The overall training procedure is as follows. The speaker verification model in Figure 2 is pre-trained on a fairly large data set particularly for the task of speaker verification. In Figure 3, this pre-trained model then initiates the same speaker model in the joint training structure. With the information of the target speaker label in the mixed signal, we thus use the enrolled speech utterances of the same target speaker as positive samples and random selected utterances of non-target speakers as negative samples for triplet sampling. The total loss defined in Eq. (6) is end-to-end back-propagated on both speaker and enhancement models, such that the speaker embedder is updated accordingly. Five enrolled positive samples are further used for speaker embedding generation with a mean pooling operator for simplicity.

4. EXPERIMENTAL SETUP

To facilitate the target speech enhancement module with robust speaker embedding on Mandarin language. External Mandarin corpora are explored to train a speaker verification model, which is then used as a pre-trained model for next stage joint training of speaker embedding and target speech enhancement. Specifically, a subset of King-ASR-216 and King-ASR-210, consisting of 5387 speakers [1], is employed for pre-training the speaker verification model. The training data is then augmented 3 folds to incorporate variabilities from distance (reverberation), channel or background noise. 40-d log mel filter bank features are extracted with a 25ms window and the time shift of feature frames is 10ms. The non-speech part is removed by a energy based voice activity detection. The utterance is randomly segmented into 200-400 frames to control the duration variability in the training phase.

[1] http://en.speechocean.com/datacenter/details/254.html



(a) Pre-trained model (b) Finetuned model (c) Joint model

Fig. 4. The visualization of speaker embeddings by (a) pre-trained speaker verification model (b) pre-trained speaker verification model and finetuned by AISHELL-2 data (c) jointly trained speaker verification model, respectively.

Table 1. *Speaker verification benchmarks (EER) on Voxceleb and AISHELL-2*

system	Voxceleb	AISHELL-2
i-vector [34]	5.33%	3.72 %
x-vector [34]	3.14%	3.21%
our model	**2.10 %**	**2.06 %**

AISHELL-2 [33], an open-sourced, self-contained baseline for industrial-scale Mandarin ASR research, is used for creating training, validation and testing data for the evaluation of the proposed approach. AISHELL-2 corpus contains 1000 hours of clean reading speech data. There are 1991 speakers participated in the recording, including 845 male and 1146 female. Age of speaker ranges from 11 to over 40. 1347 speakers are recorded in a studio, while the rest are in a living room with natural reverberation. For each speaker (as a target speaker), we randomly select 50 recorded far-field utterances of this speaker and mix each of them with utterances from other speakers at a randomly sampled signal-to-interference ratio (SIR) in {-6dB, 0dB, 6dB}. The number of speakers in the mixed signal ranges from 1 to 3 with equal chance. For 1-speaker case, the SIR is inf. An environmental noise is added to the speech mixture with signal-to-noise ratio (SNR) randomly sampled from {6dB, 12dB, 18dB, 24dB, 30dB} with equal likelihood. Noise types include kitchen, TV, home appliance, music and other ambient noises sampled from recordings of "daily life" environments. We generate 1900 speakers $\times$ 50 utterances, 60 speakers $\times$ 50 utterances, and 30 speakers $\times$ 50 utterances for training, validation and testing, respectively. The speakers in training, validation and testing sets are not overlapped.

5. RESULTS AND DISCUSSION

We first validate the the speaker verification model in Fig. 2 by two corpora, Voxceleb and AISHELL-2. Voxceleb is currently one of the most popular benchmarks. For fair comparison with standard Kaldi x-vector recipe, we adopt all the feature processing parts from Kaldi recipe except for the network and the training objective. Since our simulated mixed speech is based on clean AISHELL-2 corpus, we also report speaker verification results on a 50-speaker subset of AISHELL-2. Table 1 lists the performance of our proposed speaker verification model against i-vector and x-vector benchmarks reported by Kaldi. We are able to achieve about +30% relative improvement over the finetuned x-vector and i-vector systems.

The experimental results on target speech enhancement are shown in Table 2 and 3, evaluated by signal-to-distortion ratio (SDR) and SI-SNR, respectively. The two pre-trained methods (mean pooling and attention) shown in Figure 1 and our proposed joint learning architecture in Figure 3 are compared in single-speaker (i.e. task of denoising), two-speaker and three-speaker test cases, respectively. Furthermore, based on the pre-trained speaker verification model, we use AISHELL-2 data to finetune it for a fair comparison with the joint learning model since the speaker embedder is updated in the joint model on AISHELL-2 data as well. The single-speaker denoising task is not challenging due to moderate to high levels of SNR. As a result, there is not much difference among these approaches. The task becomes more challenging as the number of speakers in the mixture increases. The attention based context-dependent model performs better than the mean pooling based fixed speaker embedding model particularly in two-speaker and three-speaker cases, while using the unified model with joint learning leads to a larger gain of about 0.8 dB in two-speaker case and nearly 1.3 dB in three-speaker case, respectively.

Table 2. *The comparison of different methods in SDR (dB)*

Method	1 spk	2 spks	3 spks	Avg.
Raw	13.17	-0.73	-1.74	3.69
Pre-train: mean pooling	21.11	8.37	6.55	12.76
Pre-train: attention	21.31	9.42	7.88	13.62
Finetune: mean pooling	21.19	9.77	7.97	13.71
Joint-train: mean pooling	**21.46**	**10.56**	**9.24**	**14.47**

Table 3. *The comparison of different methods in SI-SNR (dB)*

Method	1 spk	2 spks	3 spks	Avg.
Raw	13.50	-0.80	-1.87	3.74
Pre-train: mean pooling	20.80	7.82	6.07	12.32
Pre-train: attention	21.00	8.88	7.38	13.17
Finetune: mean pooling	20.89	9.17	7.44	13.25
Joint-train: mean pooling	**21.17**	**9.98**	**8.66**	**14.00**

Furthermore, to evaluate the discriminability of the speaker embedding obtained through pre-trained & finetuned speaker models and jointly adapted speaker model, we use t-Distributed Stochastic Neighbor Embedding (t-SNE) [35] to visualize high-dimensional speaker representations. The utterance-level speaker embeddings from 10 speakers (4 females and 6 males), each with 50 utterances, are plotted in Figure 4. The ratio of inner-class cosine distance over inter-class cosine distance for three speaker embedders, pre-trained model, finetuned model and jointly trained model, are 0.60, 0.49 and 0.42, respectively, showing that the jointly updated system learns to make tighter speaker cluster and enlarge the separateness between speakers.

6. CONCLUSIONS

We proposed an joint learning architecture of time-domain target speech enhancement and speaker embedding model in this paper. The jointly optimized speaker embedding shows its merit over the conventional pre-trained speaker verification model for speaker embedding extraction and thus improves the target speech enhancement performance by nearly 1 dB in SDR and SI-SNR, respectively. We also show that the gain becomes more significant with the increase of the number of speakers in the mixture. The speaker embeddings extracted from the jointly optimized model become more discriminative. As a result, our future work will be towards robust speaker verification by multi-task learning on target speaker enhancement.

7. REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 30–42, 2012.

[3] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *the Proceedings of ICASSP*, 2016, pp. 31–35.

[4] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation." in *the Proceedings of ICASSP*, 2017, pp. 246–250.

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

[6] M. Yu, X. Ji, Y. Gao, L. Chen, J. Chen, J. Zheng, D. Su, and D. Yu, "Text-dependent speech enhancement for small-footprint robust keyword detection," in *Interspeech*, 2018.

[7] Z.-Q. Wang and D.-L. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM TASLP*, vol. 27, pp. 457–468, 2019.

[8] R. Gu, L. Chen, S. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Neural spatial filter: Target speaker speech separation assisted with directional information," in *Interspeech*, 2019.

[9] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, 2018.

[10] J. Wu, Y. Xu, S. Zhang, L. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," *arXiv:1904.03760*, 2019.

[11] J. Wang, J. Chen, D. Su, L. Chen, M. Yu, Y. Qian, and D. Yu, "Deep extractor network for target speaker recovery from single channel speech mixtures," in *Interspeech*, 2018.

[12] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. Saurous, R. Weiss, Y. Jia, and I. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," *arXiv preprint arXiv:1810.04826*, 2018.

[13] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan, C. Liu, D. Dimitriadis, J. Droppo, and Y. Gong, "Single-channel speech extraction using speaker inventory and attention network."

[14] Z.-Q. Wang, J. L. Roux, D. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," *arXiv preprint arXiv:1804.10204*, 2018.

[15] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 483–492, 2016.

[16] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *the Proceedings of ICASSP*, 2018.

[17] ——, "Tasnet: Surpassing ideal time-frequency masking for speech separation," *arXiv preprint arXiv:1809.07454*, 2018.

[18] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," in *ISCA INTERSPEECH*, 2017.

[19] C. Zhang, K. Koishida, and J. H. L. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM TASLP*, vol. 26, no. 9, pp. 1633–1644, 2018.

[20] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE ICASSP*, 2018.

[21] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin *et al.*, "State-of-the-art speaker recognition for telephone and video speech: the jhu-mit submission for nist sre18," in *ISCA INTERSPEECH*, 2019.

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

[23] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *ICASSP*, 2018.

[24] R. Li, N. Li, D. Tuo, M. Yu, D. Su, and D. Yu, "Boundary discriminative large margin cosine loss for text-independent speaker verification," in *ICASSP*. IEEE, 2019, pp. 6321–6325.

[25] C. Zhang, F. Bahmaninezhad, S. Ranjan, H. Dubey, W. Xia, and J. H. Hansen, "Utd-crss systems for 2018 nist speaker recognition evaluation," in *the Proceedings of ICASSP*, 2019.

[26] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of CVPR*, 2018, pp. 5265–5274.

[27] Y. Liu, L. He, and J. Liu, "Large margin softmax loss for speaker verification," in *ISCA INTERSPEECH*, 2019.

[28] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of CVPR*, 2015, pp. 815–823.

[29] Z. Chen, S. Watanabe, H. Erdogan, and J. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Interspeech*, 2015.

[30] X. Li and X. Wu, "Modeling speaker variability using long short-term memory networks for speech recognition," in *Interspeech*, 2015.

[31] S. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-end attention based text-dependent speaker verification," in *SLT*, 2016.

[32] S. Shon, H. Tang, and J. Glass, "Voiceid loss: Speech enhancement for speaker verification," in *Interspeech*, 2019.

[33] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv:1808.10583*, 2018.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *ASRU*, 2011.

[35] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.