INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Target Speaker Extraction for Multi-Talker Speaker Verification

Wei Rao[1], *Chenglin Xu*[2,3], *Eng Siong Chng*[2,3], *Haizhou Li*[1]

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]Temasek Laboratories, Nanyang Technological University, Singapore

elerw@nus.edu.sg, xuchenglin@ntu.edu.sg

Abstract

The performance of speaker verification degrades significantly when the test speech is corrupted by interference from non-target speakers. Speaker diarization separates speakers well only if the speakers are not overlapped. However, if multiple talkers speak at the same time, we need a technique to separate the speech in the spectral domain. In this paper, we study a way to extract the target speaker's speech from an overlapped multi-talker speech. Specifically, given some reference speech samples from the target speaker, the target speaker's speech is firstly extracted from the overlapped multi-talker speech, then the extracted speech is processed in the speaker verification system. Experimental results show that the proposed approach significantly improves the performance of overlapped multi-talker speaker verification and achieves 64.4% relative EER reduction over the zero-effort baseline.

Index Terms: target speaker extraction, overlapped speech, speaker verification.

1. Introduction

The performance of speaker verification degrades significantly when the speech is corrupted by interference speakers. Speaker diarization can be useful for speaker verification with non-overlapping multi-talker speech [1–6]. It can effectively exclude unwanted speech segments when the speakers only slightly overlap [7, 8]. However, such system fails when multi-talkers speak simultaneously most of the time.

One possible solution is to separate the multi-talker speech into different speakers using a speech separation system, such as deep clustering [9], deep attractor network [10], permutation invariant training [11–13], and so on. While speech separation performance has been improved recently, the number of speakers has to be known in prior for these approaches to work well. However, the number of speakers in the test speech is unknown in many real-world applications.

In speaker verification, what we are interested in is the target speaker's speech. We are not interested in speech of other speakers. The speaker extraction mechanism requires some clean speech samples from the target speaker as the reference [14–18]. In the context of speaker verification, the target speaker's enrollment speech can be used as such reference. We can construct a multi-talker speaker verification system by using a target speaker extraction front-end, that is followed by a traditional speaker verification system such as i-vector/PLDA system [19–22]. We call the processing pipeline as SE-SV.

In this paper, we explore the interaction between the speaker verification system and the target speaker extraction networks. The speaker extraction networks are SBF-MTSAL [16] and SBF-MTSAL-Concat [16]. Experimental results suggest that SE-SV significantly improves the performance of speaker verification on overlapped multi-talker speech over both zero-effort system and the oracle speaker diarization system.

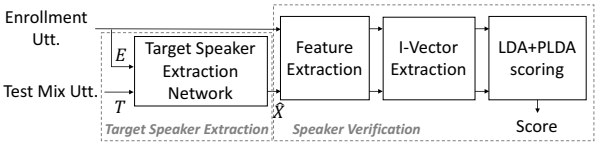

Figure 1: *The flow chart of overlapped multi-talker speaker verification system with target speaker extraction. "T" represents the mixture speech. "E" represents the enrollment speech. "$\hat{X}$" represents the extracted target speaker's speech from T.*

The remainder of the paper is organized as follows. Section 2 introduces our proposed SE-SV framework in this paper. Section 3 and Section 4 report the experimental setup and results. Then, the conclusions and future works are presented in Section 5.

2. Multi-Talker Speaker Verification with Speaker Extraction

Fig. 1 illustrates the framework of the proposed overlapped multi-talker speaker verification system with target speaker extraction (SE-SV). The framework consists of a target speaker extraction module and a speaker verification system. Specifically, the enrollment speech E is used as the reference sample for the speaker extractor to extract target speech from the mixture T. The extracted speech $\hat{X}$ instead of the original test multi-talker mixture T is used to extract i-vector for speaker verification.

In this paper, we compare the use of two target speaker extraction methods and their interaction with speaker verification system: (1) SpeakerBeam front-end with magnitude and temporal spectrum approximation loss (SBF-MTSAL) [16] and (2) SBF-MTSAL with concatenation framework (SBF-MTSAL-Concat) [16]. Both methods are the extensions to SpeakerBeam front-end (SBF) [15].

2.1. SBF-MTSAL

Fig. 2 shows the architecture of SBF-MTSAL [16]. The SBF-MTSAL framework consists of a speaker extraction network and an auxiliary network. The auxiliary network learns adaptation weights from the target speaker's voice E, which is different from the utterance of the target speaker in the mixture. The adaptation weights contain speaker characteristics. They are used to weight the sub-layers in the adaptation layer of the mask estimation network with a context adaptive deep neural network (CADNN) structure [23].

Instead of optimizing the network for an objective loss between ideal binary mask and the estimated mask [15], SBF-

http://dx.doi.org/10.21437/Interspeech.2019-1410

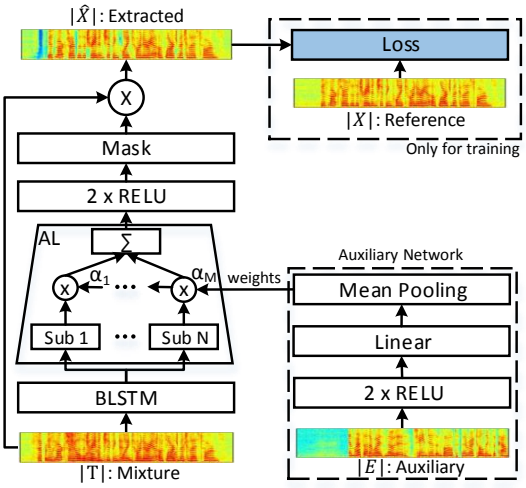

Figure 2: *The architecture of SBF-MTSAL. "AL" in the trapezium box represents the adaptation layer. "Sub" represents the sub-layer. "α" represents the weight obtained from the auxiliary network. "N" represents the number of sub-layers. "$|T|$: Mixture" represents the magnitude of the mixture speech. "$|\hat{X}|$: Extracted" represents the output magnitude of the extracted target speaker's speech. "$|X|$: Reference" represents the magnitude of clean speech, which is used to simulate the mixture. "$|E|$: Auxiliary" represents the magnitude of the enrolled target speaker's speech. During the evaluation, the upper right dotted box is not necessary.*

MTSAL computes a magnitude and temporal spectrum approximation loss to estimate a phase sensitive mask [24] due to its better performance [12, 16]. The magnitude and its dynamic information (i.e., delta and acceleration) are used in calculating the objective loss for temporal continuity. The objective loss is defined as,

$$\begin{aligned} J = \frac{1}{N_T}\sum(||&|\hat{X}| - |X| \odot cos(\theta_y - \theta_x)||_F^2 \\ &+ w_d||f_d(|\hat{X}|) - f_d(|X| \odot cos(\theta_y - \theta_x))||_F^2 \\ &+ w_a||f_a(|\hat{X}|) - f_a(|X| \odot cos(\theta_y - \theta_x))||_F^2) \end{aligned} \quad (1)$$

where the extracted speech $|\hat{X}|$ is equal to $M \odot |T|$, and M is the estimated phase sensitive mask for target speaker. $|T|$ and $|X|$ are the magnitudes of the mixture and the target speaker's clean speech, where θ_y and θ_x are their corresponding phase angles. w_d and w_a are the weights, which are tuned as 4.5 and 10.0 in this work. $f_d(\cdot)$ and $f_a(\cdot)$ are the delta and acceleration computation functions [25].

With the magnitude and temporal spectrum approximation loss, the mask estimation network and the auxiliary network obtaining the adaptive weights are jointly optimized.

2.2. SBF-MTSAL-Concat

Fig. 3 illustrates the architecture of SBF-MTSAL-Concat framework [16] that extracts target speaker's speech from the mixture. Different from the adaptive weights learned from frame level features by the auxiliary network in the SBF-MTSAL, the auxiliary network in SBF-MTSAL-Concat learns a speaker embedding from a different utterance of target speaker by a bidirectional long short-term memory networks (BLSTM). The BLSTM learns contextual information from the history and future frames of a whole utterance. The D dimensional speaker

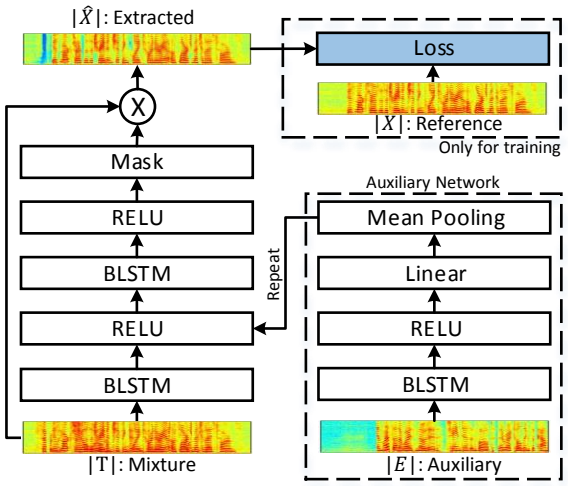

Figure 3: *The architecture of SBF-MTSAL-Concat. "$|T|$, $|\hat{X}|$, $|X|$, $|E|$" are the same annotations as in Fig. 2. During the evaluation, the upper right dotted box is not necessary.*

embedding ($V \in R^D$) is obtained by mean pooling the learnt contextual information over all frames.

Then, the speaker embedding representing speaker characteristics is repeatedly concatenated with the activation of another BLSTM in the mask estimation network. The concatenated representations containing the mixture and target speaker information are used to estimate a phase sensitive mask with the same loss function as defined in equation (1). The loss is minimized to jointly optimize the mask estimation network and the auxiliary network that learns speaker embedding.

3. Experimental Setup

3.1. Speech Data

WSJ0 corpus [26] was used to simulate the two-speaker mixed database for target speaker extraction and speaker verification experiments. We selected the utterances from WSJ0 corpus by following [9]. These utterances were considered as *Clean Dataset* and divided into three sets: training, development, and evaluation. Specifically, 11,560 utterances from 50 male and 51 female speakers were selected from WSJ0 "si_tr_s" set as the training and development sets. Among these utterances, 8,769 utterances were used as training set and another 2,791 utterances were selected as development set. The evaluation set included 1,857 enrollment utterances and 1,478 test utterances from 10 male and 8 female speakers in the WSJ0 "si_dt_05" and "si_et_05" sets. We down-sampled the database to 8kHz. The utterances in the database have an average duration of 7 seconds.

We used the *Clean Dataset* to generate the *Two-speaker Mixed Dataset*[1], which also consisted of training, development, and evaluation sets. Each set was generated by the corresponding clean dataset. Specifically, the training set of two-speaker mixed dataset included 20,000 mixtures which were generated by randomly mixing the utterances in the training set of clean dataset; Similarly, 5,000 and 3,000 mixed utterances were generated as the development set and test utterances in the evaluation set for the two speaker mixed database, respectively. Considering the interference speech as noise, the SNR of each

[1]The database simulation code is available at: `https://github.com/xuchenglin28/speaker_extraction`

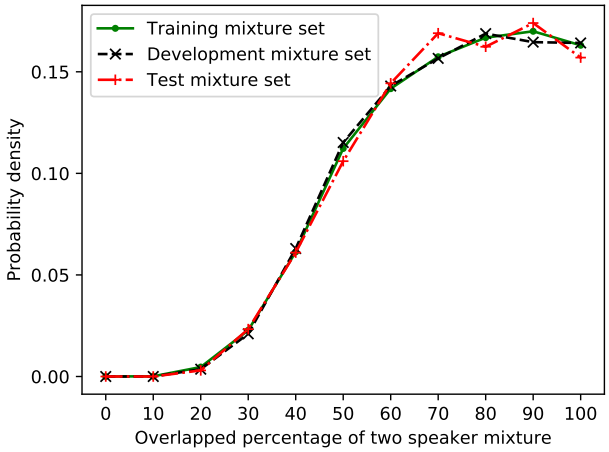

Figure 4: *The overlapped percentage of two speaker mixture on training set, development set, and test set for evaluation in the two-speaker mixed dataset.*

mixture was randomly selected between 0dB and 5dB. The enrollment utterances in the evaluation set of two-speaker mixed database were kept same as the those in the clean dataset.

In the simulation of two-speaker mixture, the first selected speaker was chosen as target speaker, the other one was interference speaker. The utterance of the target speaker from the original WSJ0 corpus was used as reference speech. Another different utterance of this target speaker was randomly selected to be used as input to the auxiliary network. The mixtures were generated based on the rule of maximum duration. For example, if the duration of utterance A is 10 seconds and that of utterance B is 5 seconds, the duration of two speaker mixture would be 10 seconds. Therefore, the overlapped percentage of this mixture utterance is 50%. Figure 4 shows the overlapped percentage of two speaker mixture on training set, development set, and test utterances of evaluation set in the two-speaker mixed database. Most of two speaker mixtures are highly overlapped with the average length of 8.5 seconds.

Since in both clean and two-speaker mixed datasets, the speakers in the evaluation set were different from the training and development sets, the evaluation set was used to evaluate the speaker verification performance. The interference speaker used to generate the test mixtures are not used as the enrollment in the speaker verification trials involving these test mixtures. The details are described in Section 3.3.

3.2. Target Speaker Extraction Network Setup

A short-time Fourier transform (STFT) was used with a window length of 32ms and a shift of 16ms to obtain the magnitude features from both of the input mixture for mask estimation network and input target speech for auxiliary network. The normalized square root hamming window was applied.

The learning rate started from 0.0005 and scaled down by 0.7 when the training loss increased on the development set. The minibatch size was set to 16. The network was trained with minimum 30 epochs and stopped when the relative loss reduction was lower than 0.01. The Adam algorithm [27] was used to optimize the network.

The aforementioned magnitude extraction configuration and network training scheme were kept same in both SBF-MTSAL and SBF-MTSAL-Concat methods. The extracted magnitude were reconstructed into time-domain signal with phase of the mixture. Then the time-domain signal was used as input to the speaker verification system.

For SBF-MTSAL, the auxiliary network was composed of 2 feed-forward relu hidden layers with 512 hidden nodes and a linear layer with 30 hidden nodes. The adaptation weights were obtained by averaging these 30 dimensional outputs over all the frames. The mask estimation network used a BLSTM with 512 cells in each forward and backward direction. The following adaptation layer had 30 sub-layers. Each sub-layer had 512 nodes with 1024 dimensional inputs from the outputs of the previous BLSTM. The 30 dimensional weights from the auxiliary network were used to weight these sub-layers, respectively. Then the activation of the adaptation layer was summed over all the sub-layers. Another 2 feed-forward relu hidden layers with 512 nodes were appended. The mask layer had 129 nodes to predict the mask for the target speaker.

For SBF-MTSAL-Concat, the auxiliary had a BLSTM with 256 cells in each forward and backward direction, a feed-forward relu hidden layer with 256 nodes and a linear layer with 30 nodes. The output of the linear layer was averaged over all frames to obtain a 30 dimensional speaker embedding containing target speaker characteristics. The speaker embedding was repeatedly concatenated with the activation of the BLSTM layer in the mask estimation network. The BLSTM had 512 cells in each forward and backward direction. Then the concatenated outputs were fed to a feed-forward relu hidden layer, a BLSTM layer and another feed-forward relu hidden layer. The BLSTM had 512 cells and the relu layers had 512 nodes. The mask layer had 129 nodes.

3.3. Speaker Verification (SV) System

In the evaluation set of the two-speaker mixed dataset, we have 3000 target trials and 48,000 non-target trials for the SV evaluation. In the evaluation trials, each enrollment utterance contained a contiguous speech segment from a single speaker and test utterance contained the overlapped speech from two speakers. We called this evaluation set as *Mixture evaluation set*. Moreover, to show the upper bound of target speaker extraction on SV, we also generated corresponding *Clean evaluation set*[2] with 51,000 trials from WSJ0 corpus according to the target speaker information of mixture set.

The training and development sets of clean dataset, called as *Clean(training&dev)*, were used for training UBM, total variability matrix, LDA, and PLDA models. Because this paper directly used the extracted target speaker's speech for SV, it would cause the mismatch between extracted speech and clean speech. To solve this mismatched problem, we pooled 2,791 extracted speech from the development set of the simulated two-speaker mixed dataset and clean training set to train SV system. We called this training set as *Clean(training)+Ext1 set*. For further investigating the effectiveness of using extracted speech to train SV system, we also pooled all 5,000 extracted speech of the development set in two-speaker mixed dataset and the training set of clean dataset to train SV system. This pooling dataset was called as *Clean(training)+Ext2 set*. Section 4 will show the performance by using different training and evaluation set.

The features of SV system were based on 19 MFCCs together with energy plus their 1st- and 2nd-derivatives extracted from the speech regions, followed by cepstral mean normalization [28] with a window size of 3 seconds. A 60-dimensional acoustic vector is extracted every 10ms, using a Hamming win-

[2]The SV evaluation trials and keys for clean and mixture evaluation sets are available at: `https://github.com/xuchenglin28/speaker_extraction`

Table 1: *Performance of SV system with and without target speaker extraction. "Training" represents the type of training data. "Eval" represents the type of evaluation test data. "TSE" represents whether or which target speaker extraction method is used. "Baseline" represents the zero-effort test case where SV system is trained with clean data and evaluated on mixture data. "Upper Bound" represents the case where clean speech data are used in both training and testing, which offers the upper bound performance. "OSD-SV" represents the case where we replace the speaker extraction network in Figure 1 with an oracle speaker diarization (OSD) system. "DCF08" represents the minimum detection cost with $P_{Target} = 0.01$. "DCF10" represents the minimum detection cost with $P_{Target} = 0.001$. The details of experimental setup can be referred to section 3.3.*

System No.	Systems	Training	Eval	TSE	EER (%)	DCF08	DCF10
1 (Baseline)	SV	Clean(training&dev)	Mixture	No	21.80	0.873	0.912
2	SV	Clean(training)+Ext1	Mixture	No	21.57	0.854	0.926
3	SV	Clean(training)+Ext2	Mixture	No	21.67	0.850	0.898
4	SE-SV	Clean(training&dev)	Mixture	SBF-MTSAL	10.87	0.766	0.867
5	SE-SV	Clean(training)+Ext1	Mixture	SBF-MTSAL	8.50	0.677	0.797
6	SE-SV	Clean(training)+Ext2	Mixture	SBF-MTSAL	8.30	0.643	0.777
7	SE-SV	Clean(training&dev)	Mixture	SBF-MTSAL-Concat	10.37	0.736	0.861
8	SE-SV	Clean(training)+Ext1	Mixture	SBF-MTSAL-Concat	7.93	0.640	0.747
9	SE-SV	Clean(training)+Ext2	Mixture	SBF-MTSAL-Concat	**7.77**	**0.631**	**0.747**
10 (Upper Bound)	SV	Clean(training&dev)	Clean	No	3.00	0.360	0.522
11	SV	Clean(training)+Ext1	Clean	No	3.07	0.366	0.526
12	SV	Clean(training)+Ext2	Clean	No	3.07	0.377	0.524
13	OSD-SV	Clean(training&dev)	Mixture	No	14.60	0.851	0.908

dowing of 25ms. An energy based voice activity detection method is used to remove silence frames. The system was based on a gender-independent UBM with 512 mixtures. The training set described in the previous paragraph was used for estimating the UBM and total variability matrix with 400 total factors. The same data set was used for estimating the LDA and Gaussian PLDA models with 150 latent variables.

4. Experimental Results

To investigate the effect of overlapped test speech on SV, we perform the SV experiments on both mixture and clean evaluation set described in Section 3.3. System 1 of Table 1 is the baseline system of SV with clean training data on mixture test set. System 10 of table 1 shows the upper bound performance (also called as ideal performance) of SE-SV for overlapped multi-talker SV. Comparison between System 1 and 10 of Table 1 shows that the performance of SV system seriously degrades when the test speech is the overlapped multi-talker speech.

Table 1 presents the performances of SV systems with and without target speaker extraction. System 1 of Table 1 is the baseline results of overlapped multi-talker SV. Systems 4 to 6 and Systems 7 to 9 of Table 1 show the performances of SE-SV with SBF-MTSAL and SBF-MTSAL-Concat on multi-talker SV. The comparison of performances among Systems 1 to 9 suggests the following findings: (1) SE-SV significantly improve the performance of multi-talker SV, specifically, the performance of multi-talker SV after applying SE-SV with SBF-MTSAL-Concat can obtain around $64.4\%, 27.7\%, 18.1\%$ relative reduction over the baseline in terms of EER, DCF08, and DCF10, respectively; (2) SE-SV with SBF-MTSAL-Concat outperforms SE-SV with SBF-MTSAL in terms of both EER and DCFs; (3) pooling the clean training set and extracted speech data is effective to alleviate the effect caused by mismatch problem between clean and extrated speech; (4) more extracted speeches for SV training could further improve the performance of SE-SV on multi-talker SV; (5) comparing System 2 with 5, and 8, we observe that most of improvement on the overlapped multi-talker SV is attributed to SE-SV. The same conclusion could be made by comparing System 3, 6, and 9.

This paper also investigates the performance of oracle speaker diarization on multi-talker SV. To this end, we first apply the energy-based VAD on the clean speech that makes up the mixture speech, then generate the diarization labels for the mixture speech according to the VAD labels of clean speech. Because WSJ0 speech data is very clean and energy-based VAD could works very well on it, we consider these diarization labels as oracle diarization labels. With the oracle speaker diarization, the average percentage of removed non-target speech frames in the mixture speeches is around 24%. System 13 in Table 1 shows the performance of multi-talker SV system with the oracle speaker diarization, which also means the best performance achieved by speaker diarization for multi-talker SV. The comparison of performance among System 1, 9, and 13 suggests that our proposed SE-SV method significantly outperforms speaker diarization in the overlapped multi-talker scenarios.

In Table 1, we also observe that the performance of System 11 and 12 with *Clean+Ext* training data slightly drops when comparing with System 10. This is expected as we include the mismatched data during training.

5. Conclusions and Future Works

This paper proposes SE-SV to improve the performance of SV on overlapped multi-talker speech and compares with the oracle speaker diarization on this task. Experimental results show that the proposed SE-SV approach could significantly improve the performance of multi-talker SV and outperform the oracle speaker diarization system. Joint training of target speaker extraction and speaker embedding network for SV will be studied in the future work.

6. Acknowledgements

This work is supported by the Neuromorphic Computing Programme under the RIE2020 Advanced Manufacturing and Engineering Programmatic Grant A1687b0033 in Singapore.

7. References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[2] O. Kudashev, S. Novoselov, K. Simonchik, and A. Kozlov, "A speaker recognition system for the SITW challenge." in *Proceedings of Interspeech*, San Francisco, California, Sep. 2016, pp. 833–837.

[3] Y. Liu, Y. Tian, L. He, and J. Liu, "Investigating various diarization algorithms for speaker in the wild (SITW) speaker recognition challenge," in *Proceedings of Interspeech*, San Francisco, California, Sep. 2016, pp. 853–857.

[4] O. Novotnỳ, P. Matejka, O. Plchot, O. Glembek, L. Burget, and J. Cernockỳ, "Analysis of speaker recognition systems in realistic scenarios of the SITW 2016 challenge." San Francisco, California, Sep. 2016.

[5] H. Ghaemmaghami, M. H. Rahman, I. Himawan, D. Dean, A. Kanagasundaram, S. Sridharan, and C. Fookes, "Speakers in the wild (SITW): The QUT speaker recognition system," in *Proceedings of Interspeech*, San Francisco, California, Sep. 2016, pp. 838–842.

[6] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proceedings of ICASSP*, Brighton, UK, May 2019.

[7] D. Charlet, C. Barras, and J.-S. Liénard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates," in *Proceedings of ICASSP*, 2013, pp. 7707–7711.

[8] S. H. Yella and H. Bourlard, "Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1688–1700, 2014.

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

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

[11] 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 Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[12] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid LSTM," in *Proceedings of ICASSP 2018*. Calgary, Alberta, Canada: IEEE, Apr. 2018.

[13] C. Xu, W. Rao, E. S. Chng, and H. Li, "A shifted delta coefficient objective for monaural speech separation using multi-task learning," in *Proceedings of Interspeech 2018*, Hyderabad, India, Sep. 2018, pp. 3479–3483.

[14] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Learning speaker representation for neural network based multichannel speaker extraction," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 8–15.

[15] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proceedings of ICASSP*, 2018, pp. 5554–5558.

[16] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in *Proceedings of ICASSP 2019*, arXiv preprint arXiv:1903.09952.

[17] 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 *Proceedings of Interspeech*, Hyderabad, India, Sep. 2018, pp. 307–311.

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

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

[20] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proceedings of Odyssey: Speaker and Language Recognition Workshop*, Brno, Czech Republic, Jun. 2010.

[21] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proceedings of 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.

[22] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proceedings of Interspeech 2011*, Florence, Italy, Aug. 2011, pp. 249–252.

[23] M. Delcroix, K. Kinoshita, C. Yu, A. Ogawa, T. Yoshioka, and T. Nakatani, "Context adaptive deep neural networks for fast acoustic model adaptation in noisy conditions," in *Proceedings of ICASSP*, 2016, pp. 5270–5274.

[24] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proceedings of ICASSP*, 2015, pp. 708–712.

[25] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[26] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[27] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, Jun. 1974.