International Journal of Speech Technology (2022) 25:261–268
https://doi.org/10.1007/s10772-022-09959-8

Exploring single channel speech separation for short-time text-dependent speaker verification

Jiangyu Han[1] · Yan Shi[1] · Yanhua Long[1] · Jiaen Liang[2]

Received: 31 January 2021 / Accepted: 31 December 2021 / Published online: 13 January 2022

Abstract

The automatic speaker verification (ASV) has recently achieved great progress. However, the performance of ASV degrades significantly when the test speech is corrupted by interference speakers, especially when multi-talkers speak at the same time. Although the target speech extraction (TSE) has also attracted increasing attention in recent years, its TSE ability is constrained by the required pre-saved anchor speech examples of the target speaker. It becomes impossible to directly use existing TSE methods to extract the desired test speech in an ASV test trial, because the speaker identity of each test speech is unknown. Therefore, based on the state-of-the-art single channel speech separation technique—Conv-TasNet, this paper aims to design a test speech extraction mechanism for building short-time text-dependent speaker verification systems. Instead of providing a pre-saved anchor speech for each training or test speaker, we extract the desired test speech from a mixture by computing the pairwise dynamic time warping between each output of Conv-TasNet and the enrollment utterance of speaker model in each test trial in the ASV task. The acoustic domain mismatch between ASV and TSE training data, the behaviors of speech separation in different stages of ASV system building, such as, the voiceprint enrollment, test and PLDA backend are all investigated in detail. Experimental results show that the proposed test speech extraction mechanism in ASV brings significant relative improvements (36.3%) in overlapped multi-talker speaker verification, benefits can be found not only in ASV test stage, but also in target speaker modeling.

Keywords Speaker verification · Text-dependent · Test speech extraction · Conv-TasNet

1 Introduction

In recent years, the automatic speaker verification (ASV) has achieved great success. However, in many real-world speech applications, the test speech in ASV tasks may be corrupted by interference speakers, then the ASV performances will be significantly degraded, especially when the utterances of the desired test speaker and interfering speakers are overlapped. Therefore, a real-time test speech separation front-end is very important and necessary for improving the ASV systems under multi-talker acoustic conditions.

Speech separation refers to the task of extracting all overlapping speech sources in a given mixed speech signal (Kolbæk et al., 2017). To separate the overlapped multi-talker speech into different individual speakers, there are many speech separation techniques in literature. For example, the traditional time-domain speech separation methods based on the independent component analysis (Jang et al., 2003), the sparse non-negative matrix factorization (Schmidt & Olsson, 2006) and the source-filter (Stark et al., 2011); The proposed time–frequency (T–F) masking methods based on deep learning, such as the deep ensemble separation (Zhang & Wang, 2016), deep clustering (Isik et al., 2016; Wang et al., 2019), deep attractor network (Chen et al., 2017), permutation invariant training (Kolbæk et al., 2017; Yousefi et al., 2019), etc. Although the automatic speech separation has achieved great success, most of these methods are difficult

✉ Yanhua Long
yanhua@shnu.edu.cn

Jiangyu Han
jyhan03@163.com

Yan Shi
shiyanilj@163.com

Jiaen Liang
liangjiaen@unisound.com

[1] Key Innovation Group of Digital Humanities Resource and Research, Shanghai Normal University, Shanghai 200234, China

[2] Unisound AI Technology Co., Ltd., Beijing, China

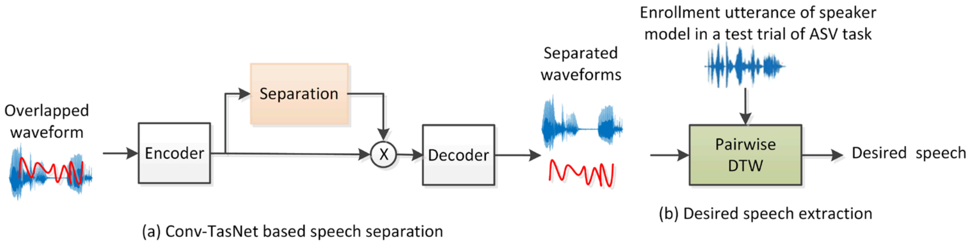

Enrollment utterance of speaker model in a test trial of ASV task
Separated waveforms
Separation
Overlapped waveform
Encoder
X
Decoder
Pairwise DTW
Desired speech
(a) Conv-TasNet based speech separation
(b) Desired speech extraction

Fig. 1 The block diagram of the proposed desired speech extraction

to meet the real-time, high accuracy requirement in ASV applications with single-channel acoustic environments. In recent years, the end-to-end architecture-Conv-TasNet proposed in Luo and Mesgarani (2018, 2019) has become the mainstream speech separation technique. Due to its competitive performance, smaller model size and shorter minimum latency, many new modifications or improvements based on this architecture have been proposed (Ge et al., 2020; Luo et al., 2020). All of these advantages make it become a good real-time speech separation front-end candidate for most real-world ASV systems.

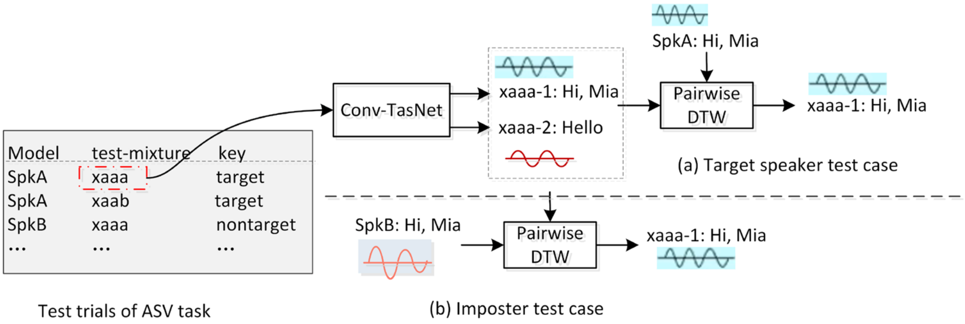
As a specific case of source speech separation, the target speech extraction (TSE), which extracts a target source in a mixture speech given clues about the target speaker, has also attracted increasing attention in recent years. Many research efforts have been paid on the TSE techniques, such as the VoiceFilter (Wang et al., 2019), SpeakerBeam (Delcroix et al., 2019; Ochiai et al., 2019), Guided source separation (Kanda et al., 2019), Speech separation using speaker inventory (Xiao et al., 2019), SBF-MTSAL based target speaker extraction (Rao et al., 2019), etc. However, most of these previous works require a pre-recorded anchor speech or guided speech of the target speaker for extracting target speaker embeddings. These embeddings are then taken as an auxiliary inputs of speech separation systems to resolve the permutation ambiguity. And for the target speech extraction based on Conv-TasNet, we also find many previous works in the literature, such as in Bahmaninezhad et al. (2019), authors used Conv-TasNet to extract target speech by introducing the multi-channel speaker location information; Wu et al. (2019) generalized Conv-TasNet to extract target speech by using audio-visual multi-modal features; The recent proposed time-domain SpeakerBeam (Delcroix et al., 2020), SpEx+ (Ge et al., 2020) and our recent works in Han et al. (2021, 2020), etc. In all of these works, a pre-recorded adaptation utterance of the target speaker is required to extract his/her voice characteristics to guide the Conv-TasNet towards extracting speech of that speaker. Although there are many previous works based on Conv-TasNet have been proposed for target speech extraction tasks, their TSE ability was still constrained by the required pre-saved anchor speech of the target speaker. This constraint makes it is impossible to directly use existing TSE methods to extract the desired test speech in an ASV test trial, because the speaker identity of each test speech is unknown and can not be provided in advance.

In this study, we focus on improving the single-channel short-time text-dependent speaker verification under multi-talker acoustic conditions. Based on the end-to-end speech separation technique-Conv-TasNet, a new test speech extraction architecture is proposed for text-dependent ASV (TD-ASV) systems. As the speaker identity of each test speech in the mixture signal is unknown and can not be provided in advance, unlike using a pre-saved anchor speech as reference in the conventional TSE tasks, here, we propose to extract the desired test speech from a mixture in a very simple but effective way. The extraction is performed by directly computing the pairwise dynamic time warping (DTW) (Berndt & Clifford, 1994; Salvador & Chan, 2007) between each output of Conv-TasNet and the enrollment utterance of speaker model in each test trial in the ASV task, no matter the speaker identity of the desired test speech is a target speaker or an imposter. The idea is motivated by the special text-dependent property and the speaker identity information matching principle of the TD-ASV tasks. The acoustic domain mismatch between ASV and speech separation system training is first investigated. Then the effects of integrating speech separation in ASV system building, target speaker enrollment and test stages are all examined and compared. Experimental results show that the proposed speech separation front-end in ASV is very effective. It brings significant relative improvements (36.3%) in overlapped multi-talker speaker verification, benefits can be found not only in ASV test stage, but also in target speaker modeling.

2 Conv-TasNet based desired speech extraction

This section presents the whole framework of the proposed Conv-TasNet based desired speech extraction for text-dependent ASV task. Figure 1 illustrates the block diagram. It consists of two main modules: (a) is a standard Conv-TasNet based speech separation that has been proposed in Luo and Mesgarani (2019), and (b) is the proposed desired speech extraction of interfered test segment in each test trial for TD-ASV. Specifically, given one two-speakers' overlapped test speech segment (mixture speech), it is firstly

Fig. 2 Example of desired test speech extraction for each test trial in ASV task



Model	test-mixture	key
SpkA	xaaa	target
SpkA	xaab	target
SpkB	xaaa	nontarget
···	···	···

Test trials of ASV task

(a) Target speaker test case

(b) Imposter test case

separated into two individual segments using the real-time Conv-TasNet speech separation system, then the desired test speech is extracted by comparing these two individual segments with the enrollment utterance of speaker model in each test trial, using the pairwise DTW criterion. As shown in Sect. 3, this framework is then taken as the desired speech extraction front-end of an ASV system.

2.1 Conv-TasNet

In speech separation, an overlapped or mixture speech $x(t)$ can be seen an additive combination of C sources $s_1(t), \dots, s_c(t)$. Conv-TasNet aims to estimate each of the C sources directly from the waveform $x(t)$. From Luo and Mesgarani (2019), we know that the Conv-TasNet is a fully-convolutional time-domain audio separation network, as shown in block (a) of Fig. 1, it consists of three modules: an encoder, a separator and a decoder.

The encoder is a 1-D convolution operation followed by a ReLu activation function. By encoding short-segments of the mixture waveform, we can get their corresponding high-dimensional representations. Then the representation is fed into the separator to estimate a multiplicative mask for each source and for each encoder output at each time step. This separator consists of a layer normalization followed by a standard convolution with kernel size 1, a series of 1-D convolution blocks, and a standard convolution block with kernel size 1 followed by a softmax operation. In each 1-D convolution block, the dilated convolution was used to capture the long-range dependencies of the speech signal. Moreover, the depthwise separable convolution was used here to reduce the number of parameters. Finally, the source waveforms are then reconstructed by transforming the masked encoder representations using the decoder with a 1-D linear deconvolution operation. More details of the Conv-TasNet architecture can be found in Luo and Mesgarani (2019).

2.2 Short-time text-dependent desired speech extraction

As stated in the introduction, this study focuses on the short-time text-dependent ASV system, our goal is to design a speech extraction front-end to extract the desired test speech segment from a mixture (interfered speech) in each test trial. However, in an ASV system, the speaker identity of each test speech in the mixture signal is needed to be predicted, it is unknown and can not be provided in advance. That's to say, it is impossible to provide any pre-saved anchor speech as a target speaker clue to leverage the Conv-TasNet to extract the target speech. Therefore, we can not use any state-of-the-art TSE methods to perform the desired speech extraction in the ASV test stage.

Unlike developing a speaker embedding extraction system to provide target speaker clues to build a TSE system, we choose to utilize the enrollment utterance of speaker model in each test trial in the ASV task to pick the desired test speech. This idea is motivated by the ASV text-dependent property and permutation invariant training criterion used in Kolbæk et al. (2017), and it also takes the advantage of the imposter speaker identity information matching principle of the TD-ASV tasks. As shown in block (b) of Fig. 1, we first compute the pairwise DTW Salvador and Chan (2007) between each Conv-TasNet output M_i and the enrollment utterance U_{model} of the speaker model in each test trial, then the segment with minimum distance[1] is taken as our desired speech M_o as follows:

$$M_o = \underset{i \in 1,\dots,N}{\arg\min} \left\{ \mathbf{DTW}\left\{ U_{model}, M_i \right\} \right\} \tag{1}$$

In addition, inspired by the objective function of Conv-TasNet, we also tried the scale-invariant source-to-noise ratio (SI-SNR) as our objective function, and we find that the results based on DTW are significantly better than those based on the SI-SNR.

[1] https://pypi.org/project/fastdtw/.

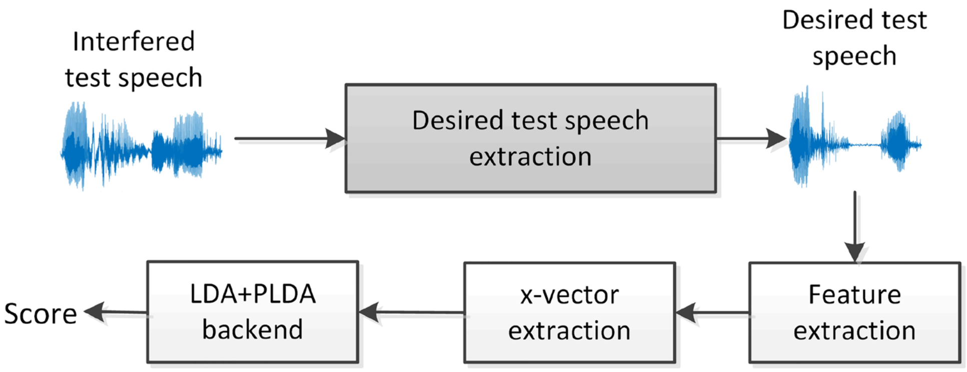


Fig. 3 Framework of the speaker verification system with a desired test speech extraction front-end

Figure 2 illustrates an example of desired test speech extraction for two test trials in the ASV task. (a) is for the target speaker test case and (b) is for the imposter test case. As shown in the list (first line) of test trials, we know that the mixture segment 'xaaa' includes a target speaker's (SpkA) speech 'xaaa-1' with the same text 'Hi, Mia' as the enrollment utterance of SpkA model. 'xaaa-2' is the interferer with text 'Hello'. After performing the pairwise DTW between 'xaaa-1' and 'SpkA', 'xaaa-2' and 'SpkA', the desired test speech can be achieved at an extremely high confidence, because both the speaker identity and context are matched with the enrollment speech of target speaker model.

However, on the contrary, as the third test trial in Fig. 2, both 'xaaa-1' and 'xaaa-2' are imposters (nontarget) for 'SpkB'. After performing the DTW matching, the achieved speech may be 'xaaa-1' or 'xaaa-2', but as 'xaaa-1' has the same context 'Hi, Mia' with its test speaker model 'SpkB', we expect that this same context constraint may lead to a better speech extraction of the desired 'xaaa-1' instead of 'xaaa-2'. In addition, even when the voiceprint matching information between 'SpkB' and 'xaaa-2' is stronger than the text-dependent context matching between 'SpkB' and 'xaaa-1', the wrongly extracted 'xaaa-2' will not have a very bad effect on the speaker verification decision of this test trial. Because for 'SpkB' model, both 'xaaa-1' and 'xaaa-2' are imposters. Therefore, we think that by utilizing the text-dependent property of TD-ASV tasks, our proposed desired test speech extraction is better than the existing target speech extraction techniques which highly depends on the pre-saved anchor reference speech. For example, if we use the state-of-the-art TSE system-TD-SpeakerBeam in Delcroix et al. (2020) to extract the test segment in the third test trial, its output may not be either 'xaaa-1' or 'xaaa-2', it may be still a mixture signal that has stronger voiceprint relationship with 'SpkB' than both of 'xaaa-1' and 'xaaa-2', because the reference speech of TD-SpeakerBeam can only be the 'SpkB' which is a wrong indicator for desired speech extraction. In addition, as far as we know, this study is the first investigation for the desired speech extraction of interfered test segment in each ASV test trial.

3 Speaker verification

Figure 3 demonstrates the framework of our speaker verification system with the proposed desired speech extraction front-end. Given an interfered test speech segment (mixture), we first extract the desired speech using the proposed front-end as shown in Fig. 1; Then the extracted speech instead of the original interfered test one is fed into the standard x-vector (Snyder et al., 2017, 2018) based speaker verification system.

As shown in the below part of Fig. 3, the standard x-vector system consists of three modules: feature extraction, x-vector extractor and the LDA plus PLDA backend (Snyder et al., 2018). In our experiments, this standard system is taken as our speaker verification baseline.

4 Experiments and results

4.1 Dataset

The HI-MIA corpus (Qin et al., 2019) was used to construct our short-time text-dependent speaker verification task. This corpus was used in AISHELL Speaker Verification Challenge 2019, and the contents are wake-up words "Hi, Mia" in Chinese. The data was collected in real home environment using microphone arrays and Hi-Fi microphone. Only the utterances recorded using Hi-Fi microphone were selected to generate our ASV task. Specifically, 40 (one per target speaker) and 2959 utterances were selected as the target speaker's enrollment and test datasets respectively. Based on these utterances, 2959 target and 115,401 non-target (imposter) test trials were generated for ASV evaluation. The rest of the 20,267 utterances from 242 speakers were selected as the development dataset to train our x-vector network and PLDA backend. There is no overlap between the enrollment, test and development datasets.

To investigate the domain mismatch between the speech separation and ASV, two datasets are constructed to simulate the two-speaker mixed conditions for Conv-TasNet model training. Besides the HI-MIA corpus, the Aishell-1 Mandarin conversational speech corpus (Bu et al., 2017) is also used. The first dataset we called "SS1", is only generated from the Aishell-1 corpus, 18,000 and 12,000 source utterances are selected and randomly mixed to generate a total of 26,991 and 5998 mixtures as the Conv-TasNet training and validation sets respectively. While in the second database we called "SS2", 15,688 source utterances are selected from the HI-MIA corpus, and 15,688 source utterances are selected from the Aishell-1. These two sources are then mixed into a total of 31,372 mixtures as the Conv-TasNet training set. For the "SS2" validation set, a total of 5918 mixtures are generated from the 2960 utterances of HI-MIA and 2960 utterances of Aishell-1. Both these two validation sets are used to tune the parameters during Conv-TasNet model training. To evaluate the speech separation performance of Conv-TasNet, we select 2959 source utterances from each of the HI-MIA and Aishell-1 corpus to generate a total of 2959 mixtures. This evaluation set is referred to "S-eval" for simplicity. All the utterances in both HI-MIA and Aishell-1 are with 16 kHz sampling rate. In addition, to simulate the interfered test utterances of ASV evaluation set, only the utterances of Aishell-1 are taken as the interfering signals. All of the data are resampled to 8 kHz from the original 16 kHz sampling rate.

4.2 Configurations

During Conv-TasNet model training, we only use global layer normalization instead of batch normalization in the separation module, all of the other configurations in this study are the same as the ones that achieved the best results in Luo and Mesgarani (2019), including the network hyperparameters, the Adam optimizer and the SI-SNR with PIT objective function.

For the ASV system training, the Kaldi main branch recipe at https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2 is used to extract the 20-dimensional MFCC plus its Δ, ΔΔ acoustic features, train the x-vector extractor and LDA plus PLDA backend. The MFCC frame-length is 25 ms with a 10 ms frame shift. Energy-based voice activity detection is applied to remove the non-speech frames. The network architecture of x-vector extractor is the same as the one presented in work (Snyder et al., 2018). Finally, the 512-dimensional x-vectors are projected to 200-dimensional by LDA, then length-normalized and modeled by PLDA.

Both the scale-invariant signal-to-noise ratio improvement (SI-SNRi) and signal-to-distortion ratio improvement (SDRi) (Vincent et al., 2006) are used to measure the speech separation performance. While for the ASV systems, the conventional equal error rate (EER) is taken as the performance measurement.

Table 1 Performance comparison on "S-eval" set between different speech separation models based on Conv-TasNet

Training	SI-SNRi (dB)	SDRi (dB)
WSJ0-2MIX	3.41	4.22
SS1	4.82	5.80
SS2	7.81	8.57

4.3 Results and discussion

4.3.1 Desired speech separation and extraction

To examine the domain mismatch behavior of Conv-TasNet based speech separation, we also reproduced the results in Luo and Mesgarani (2019) on WSJ0-2MIX corpus. Compared with the 14.6dB SI-SNRi and 15.0 dB SDRi listed in Table III of Luo and Mesgarani (2019), we obtain the corresponding 15.4 dB SI-SNRi and 15.6 dB SDRi using the same experimental setup. Besides the TasNet trained on "SS1" and "SS2" datasets, the TasNet trained on the WSJ0-2MIX is also used to separate the "S-eval" test set.

From the results shown in Table 1, it is clear to see that the Conv-TasNet model trained on SS2 achieved much better performances than the models trained on the other two datasets. It indicates that the separation performances are significantly degraded due to the domain mismatch between the Conv-TasNet training and evaluation sets. That's to say, the mismatch between conversation and wake-up speech is very big. Moreover, it is interesting to see that the Conv-TasNet trained on WSJ0-2MIX and SS1 achieved similar results, although there is big language mismatch between these two datasets, as the WSJ0-2MIX is a pure English corpus while SS1 is a pure Mandarin corpus.

To evaluate the performances of desired speech extraction algorithm for real ASV tasks, we randomly select 200 mixtures from the interfered ASV test set, 100 for target speaker test trials and the other 100 for imposter test trials. These mixtures are then taken as the evaluation set ("T-eval") for desired test speech extraction. The best Cov-TasNet model (TasNet-SS2) in Table 1 is then used for the desired test speech extraction. As shown in Figs. 1 and 2, the enrollment speech of target speaker model in each test trial is taken to compute the pairwise DTW distance between two outputs of the Conv-TasNet to extract the desired test speech segment. However, in order to see the upper bound performance of the desired speech extraction, we also tried to use the ground-truth source speech of the mixtures as references. Results with SI-SNR and DTW desired speech extraction criterions are shown in Tables 2 and 3, respectively.

In Tables 2 and 3, we calculate the accuracies for both the mixtures with ASV target and imposter test trials separately. Because in this stage, we focus more on the accuracy of

Table 2 Accuracy(%) of desired speech extraction on "T-eval" set based on TasNet-SS2 using the SI-SNR criterion

Reference speech	Target trials	Imposter trials
Ground-truth	99	100
Target speaker's enrollment	69	65

Table 3 Accuracy(%) of desired speech extraction on "T-eval" set based on TasNet-SS2 using the pairwise DTW

Reference speech	Target trials	Imposter trials
Ground-truth	99	98
Target speaker's enrollment	89	80

target speaker's speech extraction in a text-dependent ASV task, even if we extracted a wrong imposter's test speech as discussed in Sect. 2.2, it will not have a great impact on an ASV system in most real-applications.

Comparing the results of Tables 2 and 3, it can be found that when the reference speech of the desired speech extraction is ground-truth source signal, we obtain almost the same performances by using the SI-SNR and DTW criterion. However, when the reference speech is the ASV target speaker's enrollment speech, the result of using DTW is significantly better than that of using SI-SNR. This is because in ASV tasks, the test utterances can not be exactly the same as the target speaker's enrollment speech. There are intra and inter-speaker variabilities between the separated speech and the reference signal. Using DTW as criterion can exploit both the voiceprint information and the text-dependent characteristics. In fact, in the real ASV applications, it is impossible to use the ground-truth source speech as the reference signal during the desired test speech extraction front-end. Because in ASV tasks, both the test utterances of target and imposter speakers can not be obtained in advance. Therefore, the DTW will be used to perform the desired speech extraction in our system.

From Table 3, we can see that there is still a 10% accuracy gap between using ground-truth and target speaker's enrollment speech as the reference signal. This may due to the intra-speaker variability between the text-dependent enrollment and test speech segments of the same target speaker.

4.3.2 Speaker verification

In this section, we examined the effects of using Conv-TasNet based speech separation and desired speech extraction in different stages of ASV system building, including the evaluation, enrollment, x-vector and PLDA training. As shown in Table 4, the ASV performances using both the ground-truth and ASV target speaker's enrollment speech as the reference signals during desired speech extraction are investigated and compared.

By comparing the results of system 1 and 2 in Table 4, it is clear to see that the ASV performance is significantly degraded when the HI-MIA test utterances are interfered by the Aishell-1 speakers. However, when we compare the results of system 3 with the baseline system 2, a 49.9% relative EER reduction is achieved. It indicates that it is very effective to use the TasNet-SS2 as a speech separation and extraction front-end during ASV system evaluation under the interfering conditions.

From the results of system 3, 4, 5, it is interesting to find that when both the test and enrollment utterances are processed by TasNet-SS2, the EER is close to when the enrollment utterances are "Raw". In addition, by combing the "Raw" and "Separated" utterances as target speaker's

Table 4 Performances of ASV systems with and without desired speech separation and extraction

System ID	References	Training	Enrollment	Evaluation	EER(%)
1 (Upper bound)	–	Raw	Raw	Raw	2.839
2 (Baseline)	–	Raw	Raw	Mixture	18.35
3	Ground-truth	Raw	Raw	Separated	9.192
4		Raw	Separated	Separated	9.699
5		Raw	Raw + separated	Separated	7.536
6		Raw + separated	Raw + Separated	Separated	6.962
7	Target speaker's	Raw	Raw	Separated	13.35
8		Raw	Separated	Separated	13.42
9	Enrollment	Raw	Raw + separated	Separated	11.86
10		Raw + separated	Raw + separated	Separated	11.69

"Reference" represents the type of reference speech used in the desired speech extraction module based on TasNet-SS2 speech separation system. "Training" represents the type of training data for building x-vector network and PLDA. "Enrollment" represents the type of ASV target speaker's enrollment data. "Evaluation" represents the data type of ASV evaluation set. "Raw" represents the original clean data. "Separated" represents the extracted data. "Mixture" represents the interfering mixture data enrollment data, a further 58.9% relative EER reduction can be obtained. Furthermore, as shown in system 5 and 6, if we augment the x-vector and PLDA "Raw" training utterances by adding their "Separated" ones, we can obtain 62.1% relative performance improvement over the baseline. That's to say, both the ASV system training and target speaker enrollment can benefit a lot from the proposed desired speech extraction front-end.

Finally, by comparing the results in system 3 to 6 with the ones in system 7 to 10, consistent improvements of ASV performances are achieved. However, we can see that there is still large performance gap between the desired speech extraction with the ground-truth and target speaker's enrollment as the reference. It indicates that better desired speech extraction can result in better ASV system performances.

5 Conclusions and future works

This paper investigates a Conv-TasNet based desired speech separation and extraction approach to improve the performance of short-time text-dependent ASV under multi-talker conditions. Experimental results show that the proposed desired speech extraction front-end can significantly improve the ASV system not only in the evaluation stage, but also in the system training and target speaker enrollment stages. Compared with the baseline, the best system with the proposed front-end can achieve a 36.3% relative EER reduction (from 18.35 to 11.69%). Better desired speech extraction methods or how to adapt the existing target speech extraction algorithms for ASV test mixture extraction will be studied in our future work.

Acknowledgements This work is supported by the National Natural Science Foundation of China (Grant No. 62071302). The authors would like to thank Beijing Shell Shell Technology Co., Ltd., Beijing, China for the open-source HI-MIA and Aishell-1 datasets.

References

Bahmaninezhad, F., Wu, J., Gu, R., Zhang, S.-X., Xu, Y., Yu, M., & Yu, D. (2019). A comprehensive study of speech separation: Spectrogram vs waveform separation. In Proc. *Interspeech,* (pp. 4574–4578).

Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. in Proc. *KDD Workshop, 10*(16), 359–370.

Bu, H., Du, J., Na, X., Wu, B., & Zheng, H. (2017). Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In Proc. *Oriental COCOSDA,* (pp. 1–5).

Chen, Z., Luo, Y., & Mesgarani, N. (2017). Deep attractor network for single-microphone speaker separation. In Proc. *ICASSP*, (pp. 246–250).

Delcroix, M., Ochiai, T., Zmolikova, K., Kinoshita, K., Tawara, N., Nakatani, T., & Araki, S. (2020). Improving speaker discrimination of target speech extraction with time-domain speakerbeam. In Proc. *ICASSP,* (pp. 691–695).

Delcroix, M., Watanabe, S., Ochiai, T., & Kinoshita, K., et al. (2019). End-to-end SpeakerBeam for single channel target speech recognition. In Proc. *Interspeech,* (pp. 451–455).

Ge, M., Xu, C., Wang, L., Chng, E.S., Dang, J., & Li, H. (2020). Spex+: A complete time domain speaker extraction network. In Proc. *Interspeech,* (pp. 1406–1410).

Han, J., Long, Y., & Liang, J. (2020). Attention-based scaling adaptation for target speech extraction. arXiv:2010.10923.

Han, J., Zhou, X., Long, Y., & Li, Y. (2021). Multi-channel target speech extraction with channel decorrelation and target speaker adaptation. In Proc. *ICASSP*.

Isik, Y., Roux, J. L., Chen, Z., Watanabe, S., & Hershey, J. R. (2016). Single-channel multi-speaker separation using deep clustering. In Proc. *Interspeech,* (pp. 545–549).

Jang, G.-J., Lee, T.-W., & Oh, Y.-H. (2003). Single-channel signal separation using time-domain basis functions. *IEEE Signal Processing Letters, 10*(6), 168–171.

Kanda, N., Boeddeker, C., Heitkaemper, J., & Fujita, Y., et al. (2019). Guided source separation meets a strong ASR Backend: Hitachi/Paderborn University joint investigation for dinner party ASR. In Proc. *Interspeech,* (pp. 1248–1251).

Kolbæk, M., Yu, D., Tan, Z.-H., & Jensen, J. (2017). Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP), 25*(10), 1901–1913.

Luo, Y., Chen, Z., & Yoshioka, T. (2020). Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. In proc. *ICASSP,* (pp. 46–50).

Luo, Y., & Mesgarani, N. (2018). TasNet: Time-domain audio separation network for real-time, single-channel speech separation. In Proc. *ICASSP,* (pp. 696–700).

Luo, Y., & Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27*(8), 1256–1266.

Ochiai, T., Delcroix, M., Kinoshita, K., Ogawa, A., & Nakatani, T. (2019). Multimodal SpeakerBeam: Single channel target speech extraction with audio-visual speaker clues. In Proc. *Interspeech,* (pp. 2718–2722).

Qin, X., Bu, H., & Li, M. (2019). HI-MIA: A far-field text-dependent speaker verification database and the baselines. arXiv:1912.01231.

Rao, W., Xu, C., Chng, E.S., & Li, H. (2019). Target speaker extraction for multi-talker speaker verification. In Proc. *Interspeech,* (pp. 1273–1277).

Salvador, S., & Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis, 11*(5), 561–580.

Schmidt, M.N., & Olsson, R.K. (2006). Single-channel speech separation using sparse non-negative matrix factorization. In Proc. *Ninth International Conference on Spoken Language Processing*.

Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpu, S. (2017). Deep neural network embeddings for text-independent speaker verification. In Proc. *Interspeech,* (pp. 999–1003).

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors: Robust DNN embeddings for speaker recognition. In Proc. *ICASSP,* (pp. 5329–5333).

Stark, M., Wohlmayr, M., & Pernkopf, F. (2011). Source-filter-based single-channel speech separation using pitch information. *IEEE Transactions on Audio, Speech and Language Processing, 19*(2), 242–255.

Vincent, E., Gribonval, R., & Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing, 14*(4), 1462–1469.

Wang, S., Naithani, G., & Virtanen, T. (2019). Low-latency deep clustering for speech separation. In Proc. *ICASSP,* (pp. 76–80).

Wang, Q., Muckenhirn, H., Wilson, K., & Sridhar, P., et al. (2019). VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking. In Proc. *Interspeech,* (pp. 2728–2732).

Wu, J., Xu, Y., Zhang, S.-X., Chen, L.-W., Yu, M., Xie, L., & Yu, D. (2019). Time domain audio visual speech separation. In Proc. *ASRU,* (pp. 667–673).

Xiao, X., Chen, Z., Yoshioka, T., Erdogan, H., & Liu, C. et al. (2019). Single-channel speech extraction using speaker inventory and attention network. In Proc. *ICASSP,* (pp. 86–90).

Yousefi, M., Khorram, S., & Hansen, J. (2019). Probabilistic permutation invariant training for speech separation. In Proc. *Interspeech,* (pp. 4604–4608).

Zhang, X.-L., & Wang, D. (2016). A deep ensemble learning method for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 24*(5), 967–977.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.