Time-Domain Target-Speaker Speech Separation With Waveform-Based Speaker Embedding

Jianshu Zhao, Shengzhou Gao, Takahiro Shinozaki

Tokyo Institute of Technology

www.ts.ip.titech.ac.jp

Abstract

Target-speaker speech separation, due to its essence in industrial applications, has been heavily researched for long by many. The key metric for qualifying a good separation algorithm still lies on the separation performance, i.e., the quality of the separated voice. In this paper, we presented a novel high-performance time-domain waveform based target-speaker speech separation architecture (WaveFilter) for this task. Unlike most previous researches which adopted Time-Frequency based approaches, WaveFilter does the job by applying Convolutional Neural Network (CNN) based feature extractors directly on the raw Time-domain audio data, for both the speech separation network and the auxiliary target-speaker feature extraction network. We achieved a 10.46 Signal to Noise Ratio (SNR) improvement on the WSJ0 2-mix dataset and a 10.44 SNR improvement on the Librispeech dataset as our final results, which is much higher than the existing approaches. Our method also achieved an 4.9 SNR improvement on the WSJ0 3-mix data. This proves the feasibility of WaveFilter on separating the target-speaker's voice from multi-speaker voice mixtures without knowing the exact number of speakers in advance, which in turn proves the readiness of our method for real-world applications.

Index Terms: target-speaker speech separation, time-domain feature extraction

1. Introduction

Robust automatic speech recognition in realistic conditions often requires speech separation. Because of the importance of this research topic, there has been a great interest recently in using deep learning approaches to solve this problem. Most of the methods first use the Short-Time Fourier Transform (STFT) to transform the mixture signal into a time-frequency domain representation, then do separation based on it. Speech separation approaches such as Deep Clustering [1] and Permutation Invariant Training (PIT) [2] predict T-F bin masks of the individual sources where the clean source spectrograms are used as the training target. The individual sources can then be recovered by multiply mask with the mixture spectrograms. In recent years, the performance of time-frequency mask methods has significantly advanced with more sophisticated mask estimation approaches [3, 4]. However, these time-frequency-mask-prediction-based methods have several shortcomings. Firstly, the number of sources in the mixture needs to be known in advance when estimating mask. Secondly, reconstruction of the phase of the clean sources is a nontrivial problem, and the erroneous estimation of the phase is enforcing an upper bound on the accuracy of the source recovery.

In most cases, we might not want to separate all components of a mixture but rather to extract the speech of one target speaker. Some related literature exists for target-speech separation. For example, in [5], the authors achieved impressive results by training two separate neural networks: a speaker recognition network produces speaker-discriminative embeddings and a spectrogram masking network that takes both noisy spectrogram and speaker embedding as input, and produces a mask. However, a solution like that is insufficient because of the non end-to-end nature of their system. In [6, 7], the authors proposed SpeakerBeam which is an end-to-end target speaker extraction neural network architecture. An adaptation layer is used to combine the auxiliary feature extraction network and the speech separation network, and both of them are being trained simultaneously. Despite it is an end-to-end system, its performance still has an upper bound because of the nontrivial phase problem.

In order to break the upper bound of conventional time-frequency magnitude mask approaches, in [8], a fully-Convolutional Time-domain Audio Separation Network (Conv-TasNet), which directly utilizes raw waveform as the network input, has been proposed. It consists of three processing stages: encoder, separation and decoder. The encoder and decoder part are to simulate the process of STFT and iSTFT while the separation part calculates a multiplicative function (i.e., a mask) for each of the target sources. However, this method is designed for separating all sources in the mixture while in a real-world environment, the exactly number of sources in the mixture is often unknown.

As there are known drawbacks of time-frequency mask approaches, we introduce a novel speaker-dependent speech separation neural network architecture that utilize raw waveforms solely as input in this paper. Instead of generating speaker embedding from speatrogram, we directly extract speaker characteristics from the speech waveform spoken by the target-speaker. This process is similar to speaker verification. In [9], the authors proposed the RawNet which is an advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. The results indicate that directly extract embeddings from raw waveform can achieve a similar or even better performance than i-vector or other handcraft-feature-based method.

We present experimental results on publicly available WSJ0-2mix and WSJ0-3mix datasets, showing that this new architecture performed better than SpeakerBeam and PIT-based blind source separation method. To verify the generalization capability of the model, we also created a new dataset from LibriSpeech which contains more speakers in training and testing sets. The results on new dataset shows that proposed model also achieves better performance than SpeakerBeam.

The rest of this paper is organized as follows. We introduce the propose model in section 2. We describe the experimental procedures in section 3. We show the experimental results and analysis in section 4 and concludes the paper in section 5.

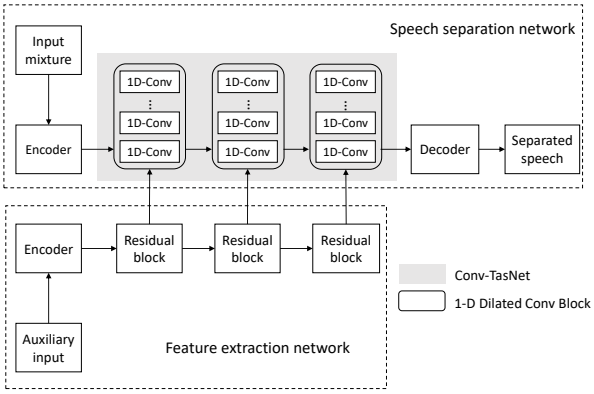


Figure 1: *Schematic diagram of proposed model*

2. Problem Definition

The problem of single-channel target speech separation is defined as estimating the target speaker source $s_t(t)$ from C speaker sources $s_1(t), \ldots, s_c(t) \in \mathbb{R}^{1\times T}$, given the mixture waveform signal $x(t) \in \mathbb{R}^{1\times T}$, where

$$x(t) = \sum_{i=1}^{C} s_i(t), \tag{1}$$

In most traditional speech separation methods, Short time Fourier transform (STFT) and inverse Short time Fourier transform (iSTFT) are used to covert time domain signal into time-frequency domain. In contrast with these conventional methods, we aim to directly estimate $s_t(t)$ from $x(t)$ in time domain.

3. Approach

The schematic diagram of the system is shown in Fig. 1. The overall system is comprised of two components: the feature extraction network and the speech separation network. The speech separation network acts as the main speech separator, and the feature extraction network extracts the target-speaker characteristic from the auxiliary waveform to aid the main separation network. Firstly, two encoders with the exact same configuration are used to transform the input mixture and the auxiliary waveform into intermediate feature spaces. The speech separation network then takes both embedding from the mixture waveform and the output of the feature extraction network as the input to estimate a mask for the target source. The separated speech is recovered by converting the masked encoder features using a decoder module. In this section, we will describe the details of each of the steps.

3.1. Feature extraction module

The purpose of the feature extraction module is to extract speaker characteristics and transform them into an intermediate space using several residual blocks. Firstly, the auxiliary speech of the target speaker is passed to the encoder which uses an 1-D convolution layer to convert the 1-dimensional signal into matrix representations. After encoding, various residual blocks are utilized to further process the encoded signal. The design of the residual block is based on the recent work in [9], which is developed for speaker verification tasks. Fig. 2 shows the details of each residual block. Each block consists of 2 CNN operations,

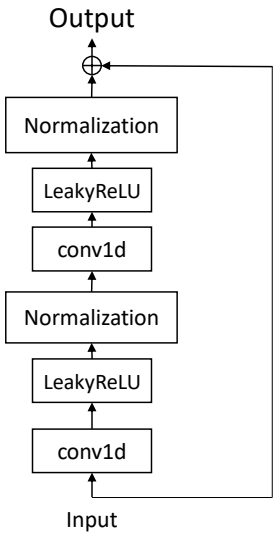


Figure 2: *Residual block*

with nonlinear activation functions and normalization added between each two convolution operations.

Motivated by the temporal convolutional network [10, 11, 12], we stack 3 residual blocks together with exponentially increasing dilation factors 1,2,4. The input to each block is zero padded accordingly to ensure the output length is the same as the input. The output of each residual block is used to adapt speech separation module to focus on the target speech in the mixture.

3.2. Speech separation module

The speech separation module is based on the related work [8] in which the authors proposed Conv-TasNet, which achieves significant performance on single channel time-domain speech separation. As shown in figure 1, the neural networks takes two input: the raw waveform of mixture signal and the speaker features computed by the residual block of the feature extraction network. The encoder structure is the same as it is in the feature extraction module.

Following the best performed configuration in Conv-TasNet, each dilated 1-D conv block consists of 8 convolution layers with nonlinear activation functions and normalization layers inserted between each two convolution operations. The output of residual block is combined with the input of the first convolution layer of each dilated 1-D conv block using an element-wise multiplication. The dilation factors increase exponentially to ensure a sufficiently large temporal context window to take advantages of the long-range dependencies of the speech signal. The target speech then reconstructed by 1-D transposed convolution operations in the decoder.

3.3. Training target

Following the recent literature on deep learning based speech separation tasks, we employ the Scale-Invariant Source-to-Noise Ratio (SI-SNR), which has commonly been used as the evaluation metric for source separation tasks, as the training cri-

terion replacing the standard source-to-distortion ratio (SDR) [13]. SI-SNR is defined as:

$$e_{noise} = \hat{s} - s_{target} \tag{2}$$

$$\text{SI-SNR} = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{noise}\|^2}, \tag{3}$$

where $s_{target} = \frac{\langle \hat{s}, s\rangle s}{\|s\|^2}$. $\hat{s} \in \mathbb{R}^{1\times T}$ and $s \in \mathbb{R}^{1\times T}$ represent the estimated and reference signals, respectively, and $\| \mathrm{s} \|^2 = \langle \mathrm{s}, \mathrm{s}\rangle$ denotes the signal power. For SI-SNR, the scale invariant is guaranteed by normalizing estimated and reference signals to zero mean.

4. Experimental Setup

In this section, we describe our experimental setup: the datasets used to train speech separation and feature extraction network also the details of two components of the system separately, as well as the metrics to assess the systems.

4.1. Data

4.1.1. WSJ0-mix data

We use the WSJ0-2mix dataset and WSJ0-3mix dataset which were introduced in [14] since it has been widely used for single-channel speech separation tasks and related works. A 30-hour-long training set and a 10-hour-long validation set of two-speaker mixtures were generated by randomly selecting utterances by different speakers from the WSJ0 training set si_tr_s, and mixing them at various SNR between 0 dB and 5 dB. The 5h test set was generated similarly using utterances from sixteen speakers from the WSJ0 development set si_dt_05 and evaluation set si_et_05. We randomly selected a clean utterance of the target speaker different from that in the mixture to be the reference utterance.

4.1.2. Librispeech 2 mixture data

To further prove the generalization capability of WaveFilter, we created a different corpus. The corpus consists of simulated mixtures of 2 speakers selected from the corpus of Librispeech [15]. We used the training and the development sets which contains 2338 speakers and 73 speakers respectively. The data in the Librispeech dataset have relatively long silences at the beginning and the end of each clip and these silences are trimmed to keep consistency with the WSJ data. As for the generation of the data, we first randomly select one amongst the 2 speakers as the target speaker and the other as the interference speaker. Then we choose an utterance from each speaker and overlay them to create the mixture input. The utterance of target speaker is the ground truth. A reference utterance of the target speaker different from the ground truth is selected as auxiliary input. According to the above method, we created a 45-hour-long training set and a 5-hour-long test set. This corpus is more challenging because it contains more speakers.

4.2. Network configuration

The inputs of the speech separation and the feature extraction network are the time-domain mixture utterance and the auxiliary utterance respectively. The utterance was encoded to an intermediate representation by using a 1-D convolution layer which has 512 filters with 32ms window length and 16ms window shift. After that, a 1×1-conv layer maps the encoded 512 dimensional features into 128 dimensions to be the input of the Conv-Tasnet. All input utterances are segmented into 4 seconds each and are resampled at 8 kHz.

We utilized the network architecture shown in Fig. 1 for all experiments. Following the configuration of the Conv-TasNet, the speech separation network consists of 3 1-D Dilated Conv Blocks, with each Dilated Conv Block being a stack of 8 1D-Conv blocks. The detailed structure of each 1D-Conv block is an exact adoption from the original paper, and can be found in [8]. For feature extraction network, it consists of 3 Residual block which has shown in Fig. 2. For the first conv1d layer in residual block it takes 128-dimension inputs and generates 512-dimension outputs. LeakyReLU nonlinear activation function with the negative slope parameter being 0.3 and a normalization layer are added between two convolution layers. The second conv1d layer takes 512-dimension inputs and generate 512-dimension outputs which are used as the inputs to each 1-D Dilated Conv Block in Conv-TasNet. The dilation and zero padding parameters of each residual block are exponentially increased from 1 to 4. The decoder consist of a conv1d transpose layer which converts the intermediate representation of the utterance back to time domain.

For training all networks, we use the Adam optimizer [16] with an initial learning rate of 0.0001 and train 200 epochs for the WSJ dataset, and 150 epochs for the Librispeech dataset. We did not use dropouts.

4.3. Evaluation

We use two metrics: Source to Distortion Ratio (SDR) and Short Term Objective Intelligibility (STOI) [17] as the evaluation criteria for our proposed model.

4.3.1. Source to distortion ratio

SDR was defined in [13] and becomes a very popular metric to evaluate the performance of source separation system. It is an energy ratio expressed in decibels (dB) and defined as:

$$\text{SDR} = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2} \tag{4}$$

where e_{interf}, e_{noise} and e_{artif} are respectively the interferences, noise, and artifacts error terms. Thus, a higher number denotes a better performance.

4.3.2. Short term objective intelligibility

Short-time objective intelligibility (STOI) is a metric that is closely related to the human auditory perception and widely used in speech separation researches as the evaluation criterion. STOI is a function of a TF-dependent intermediate intelligibility measure, which is based on the correlation between temporal envelopes of the clean and degraded speech in short-time (382 ms) segments. It has been commenly used as metrics in many related works. For example, speech enhancement [18], speech separation [19], as well as speech dereverberation and denoising [20].

5. Results

5.1. WSJ0 2 mixture data

Table 1 presented the performance comparison of WaveFilter with some of the top performing baseline models on the WSJ0 2-mix dataset. We ran the experiments on the exact same dataset

Table 1: *SDR and STOI results for WSJ0-2mix data*

	ΔSDR	**ΔSTOI**
PIT	8.6	0.101
SpeakerBeam FA 10	9.4	0.106
SpeakerBeam FA 20	9.6	0.107
SpeakerBeam FA 30	9.7	0.110
SpeakerBeam SA	9.6	0.132
WaveFilter	**10.46**	**0.152**

Table 2: *SDR and STOI results for Librispeech-2mix data*

	ΔSDR	**ΔSTOI**
SpeakerBeam SA	9.23	0.138
WaveFilter	**10.44**	**0.164**

as the baseline models, which are presented in [6] and [2], to have a fair comparison. We can see there is a clear advantage of WaveFilter over the baseline models in both SDR improvement and STOI improvement.

5.2. Librispeech 2 mixture data

To further enhance the credibility of our results, we also ran experiments for both the SpeakerBeam baseline model and WaveFilter on the Librispeech 2-mixture data for performance comparison. Table 2 presented the results. We once more observed a significant improvement in both SDR and STOI metrics for WaveFilter. However do take note that this result for the SpeakerBeam model is not written in the original paper, instead that is the result of our own reproduction of their experiment on the Librispeech 2-mixture dataset.

5.3. WSJ0 3 mixture data

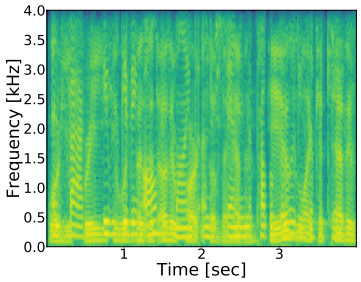
We also ran experiments using WaveFilter on the WSJ0 3-mixture dataset. The result presented in Table 3 is obtained by feeding no prior knowledge of the exact number of speakers into WaveFilter. We do see a decent SDR and STOI improvement of the resultant separated speech, and this demonstrated WaveFilter's capability in target-speaker voice extraction in a multi-speaker environment.

5.4. Example data

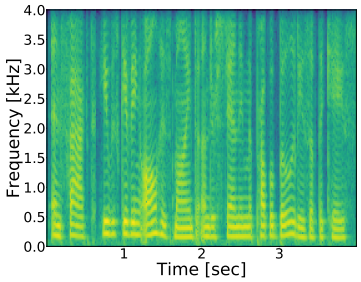
To better illustrate the effectiveness of WaveFilter, We picked a set of audio from our experimental results on WSJ0 2-mix dataset in Fig. 3 as examples. Fig. 3a is the spectrum of the mixture input for our separation algorithm, Fig. 3b is of the taget-speaker's original voice and Fig. 3c is of the extracted voice. We can see a clear distinction between the mixture clip and the extracted clip, and a noticeable similarity between the original clip and the extracted clip, which proves that WaveFilter succeeded in recovering the target-speaker's original voice from the mixture to a large extent.

Table 3: *SDR and STOI results for WSJ0-3mix data*

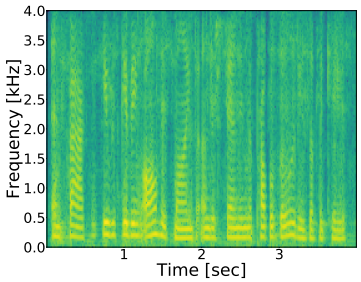
	ΔSDR	**ΔSTOI**
WaveFilter	**4.9**	**0.135**



(a) *mixture*



(b) *target-speaker original*



(c) *target-speaker extracted*

Figure 3: *Example data spectrum*

6. Conclusions and Future Work

In this paper, we presented a new model for the task of target-speaker voice separation from a voice mixture. The model is proven by experiments to be more effective than any existing solutions in the field. The full time-domain-based nature gives WaveFilter a higher degree in the end-to-endness, and thereby increase the upper bound of the performance as it makes fewer assumptions. The multi-speaker mixture compatibility broadens the scope of possible applications of WaveFilter in real-world usages.

Of course there is still room for improvements. One possible future work might be to improve the separation performance on mixture with more than 3 speakers as it can be seen that the performance on the 3-mixture dataset is relatively low under the current setting, to get WaveFilter more prepared for industrial mass implementation.

7. References

[1] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[2] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[3] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.

[4] 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.

[5] 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.

[6] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for speakerbeam target speaker extraction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6965–6969.

[7] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures." in *Interspeech*, 2017, pp. 2655–2659.

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

[9] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *arXiv preprint arXiv:1904.08104*, 2019.

[10] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 47–54.

[11] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.

[12] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[13] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[14] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

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

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[18] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[19] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[20] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 7, pp. 1492–1501, 2017.