Multi-speaker Speech Separation under Reverberation Conditions Using Conv-Tasnet

Chunxi Wang, Maoshen Jia *, Yanyan Zhang, and Lu Li

Beijing Key Laboratory of Computational Intelligence and Intelligent System, Faculty of Information Technology
Beijing University of Technology, Beijing, China; Email: chunxiwang@emails.bjut.edu.cn (C.W.),
13811321209@139.com (Y.Z.), lilubjut@163.com (L.L.)
*Correspondence: jiamaoshen@bjut.edu.cn (M.J.)

***Abstract*—The goal of speech separation is to separate the target signal from the background interference. With the rapid development of artificial intelligence, speech separation technology combined with deep learning has received more attention as well as a lot of progress. However, in the "cocktail party problem", it is still a challenge to achieve speech separation under reverberant conditions. In order to solve this problem, a model combining the Weighted Prediction Error (WPE) method and a fully-convolutional time-domain audio separation network (Conv-Tasnet) is proposed in this paper. The model target on separating multi-channel signals after dereverberation without prior knowledge of the second field environment. Subjective and objective evaluation results show that the proposed method outperforms existing methods in the speech separation tasks in reverberant and anechoic environments.**

***Keywords*—speech separation, deep learning, dereverberation, speech enhancement**

I. Introduction

Language is an important medium for communication. Separating useful information from complex scenes is a basic skill for humans. However, researchers have found in practice that it is challenging to build a speech separation system for multi-speaker that is comparable to the human auditory system [1]. Especially when the acquired signal is limited by the recording environment (e.g., the poor sound absorption performance of the wall surface and the noise level in the room is relatively high.) In recent years, deep neural networks have been introduced in speech separation tasks.

A deep attractor network (DANet) was introduced by creating attractor points in the high-dimensional embedding space of the acoustic signal. Find the centroids of the sources in the embedding space to determine the similarity of each bin in the mixture to each source [2]. To optimize the network structure, a dual-path recurrent neural network (DPRNN) approach was proposed. This model can divide longer sequences into smaller chunks, iteratively executing local and global modeling [3]. To further fuse information of adjacent phases, a Fully Recurrent Convolutional Neural Network (FRCNN) was proposed which balances the efficiency and accuracy of the model with fewer parameters [4]. In order to improve the intelligibility of speech separation, a Deep Neural Network (DNN) based speech separation method was proposed, which reduced distortion constraints [5].

Deep Learning-based speech separation systems have achieved excellent performance for pure speech separation, unfortunately its performance degrades in the reverberation environment. When there are various types of interference or noise components in the recording environment. For example, if there are two people talking at the same time in a room, the reflected and direct components of the multi-source speech signal are recorded by the microphone simultaneously [6]. The reflect components can be modeled from the sound source and room impact response [7], resulting in the destruction of the time and frequency domain sound spectrum of the original signal. With the increase of reverberation time, the reflection component increases, the interference becomes stronger, and the performance of the separation system will be dramatically reduced.

This is due to the poor separation performance caused by the lack of training in the specific acoustic environment. To achieve good results, various types of data under different reverberations are needed, and this acquisition process is complex and very costly (time and computational cost) to retrain a network model.

Therefore, this paper proposes a speech separation method under reverberation conditions by combining dereverberation and deep learning. The flowchart of the proposed method is illustrated in Fig. 1. The input mixture signals are transformed into the frequency domain by short-time Fourier transform (STFT), where the late reverberant components of the mixture signal are removed by linear prediction estimation using the Weighted Prediction Error (WPE) method [8]. The dereverberation signals are then transformed into the time domain by the inverse STFT. The signals of each channel are summed. The accumulated sum signal is feed in a fully-convolutional time-domain audio separation network (Conv-Tasnet), which employed an end-to-end deep

Manuscript received January 1, 2023; revised February 5, 2023; accepted March 22, 2023; published July 26, 2023.

learning structure. Specifically, the mixture signal is trained by this model, and the weights of the individual speakers in the mixture signal are estimated. Separated signals are obtained using the estimated weights and encoded high-dimensional representations.

Compared with existing methods, the proposed method has significant improved separation performance in reverberant conditions.

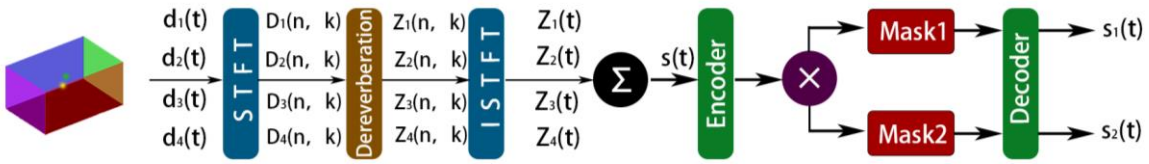

Figure 1. System flowchart.

The rest of the paper can be summarized as follows. In section II, a description of the dereverberation using the WPE method is presented. In Section III, the speech separation model based on Conv-Tasnet is described. In Section IV, the experimental procedure is described to show the experimental results and finally, Section V draws the conclusion.

II. WPE-Based Dereverberation

A. Reverberation Generation

The audio signal emitted by sound source in a confined room, reflected by different obstacles (each reflection will be absorbed by the obstacle part of the energy). After several reflections and attenuation, the reverberation signal $d_q(t)$ is obtained, which consists of both direct and reflected sound. The time in which the energy density decreases by 60dB after the source stops sounding is called the reverberation time (T60), which is used to describe the intensity of the reverberation [9].

The recorded signal $d_q(t)$ can be expressed as the sum of the convolution of the original speech source and the room impulse response:

$$d_q(t) = \sum_{i=1}^{C} s_i(t) \times r_{iq}(t) \quad (1)$$

where $s_i(t)$ is the ith speech signal and $r_{iq}(t)$ is the room impulse response (RIR) between the ith speech signal and the qth microphone. $i = 1,2,\dots,C$, C represents the number of sources. $q = 1,2,\dots,Q$, Q represents the number of microphone channels. '×' is convolution.

The reverberant component differs from the echo is the strong correlation between the direct and reflected components. Therefore, it causes confusion in the time domain and has a significant impact on speech separation.

B. Weighted Prediction Error Method

The WPE method is based on multi-channel linear prediction, which is aimed at dereverberation multi-channel recorded signals without prior knowledge [10].

The reverberation signal $d_q(t)$ received by the microphone can be divided into three parts: direct signal, early reflections, and late reflections [11]. The composition of the reverberation signal is illustrated in Fig. 2.

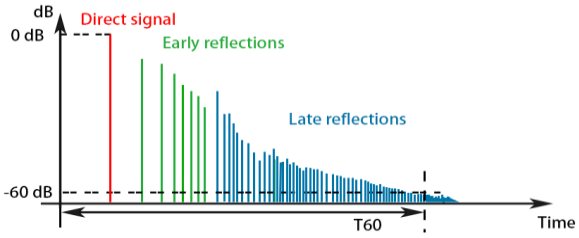

Figure 2. Composition of the reverberation signal.

The WPE method mainly removes the late reverberation component. The signal after such an operation becomes the desired dereverberation signal $z_q(t)$, which is composed of the direct signal and early reflections, which can be obtained by the following Eq. (2):

$$z_q(t) = d_q(t) - h_q(t) \quad (2)$$

where $h_q(t)$ are the late reflections that can be obtained by the Maximum Likelihood Estimate (MLE). $h_q(t)$ is expressed in the time domain as:

$$h_q(t) = \sum_{\tau=L_p}^{L_p+P-1} \sum_{q'=1}^{Q} \alpha_{q,q'}(\tau)\, d_{q'}(t-\tau) \quad (3)$$

where α represents the filter weights, the prediction step $L_p > 0$, and P is the prediction order. q and q' are the indices of different microphone channels in the array, where $q \neq q'$. Q represents the number of microphone channels. The diagram of WPE is shown in Fig. 3.

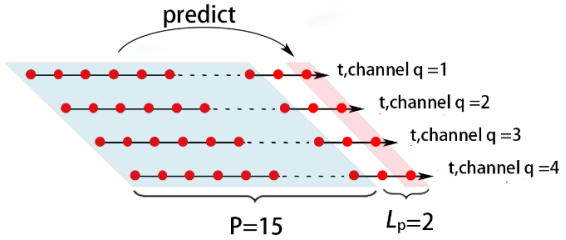

Figure 3. Diagram of WPE method.

The WPE method uses previous frames to predict the reverberation of the current frame, and then subtracts the estimated late reverberation component from the signal recorded by the microphone. The desired dereverberation signal is obtained as well as the optimal estimation of the

reverberation component in the Maximum Likelihood Estimate.

The specific process is to transform the recorded signal to the time-frequency domain using STFT, considering the sparsity of the frequency domain of the speech signal. The desired dereverberation signal $Z_q(n,k)$ in the time-frequency domain can be expressed as:

$$Z_q(n,k) = D_q(n,k) - A_q(k)D_q(n-n',k) \quad (4)$$

where $n = 1,2,\dots,N$ represents the index of frame, N denotes the total number of frames. $k = 1,2,\dots,K$ represents the index of frequency, K denotes the total number of sampling points in each frame. $D_q(n,k)$ represents the time-frequency representation of the reverberation signal. $A_q(k)$ represents the frequency domain representation of the filter weights.

And then, the desired dereverberation signal is assumed to be a time-varying Gaussian process model [12]. So, the probability density function of the desired dereverberation signal is modeled as follows:

$$p\left(Z_q(n,k);\psi_q(n,k)\right) = G_P(Z_q(n,k);0,\psi_q(n,k)) \quad (5)$$

where $G_P(\cdot)$ is the probability density function of a complex Gaussian random process with zero mean and variance $\psi_q(n,k)$.

For each frequency bin, the weights α and variances ψ take different values. The desired dereverberation signal $Z_q(n,k)$ of each time-frequency is statistically independent.

Thus, the MLE function of the model is given as follows:

$$\mathcal{L}(\alpha,\psi) = \prod_{n=1}^{N} G_P(Z_q(n,k);0,\psi_q(n,k)) \quad (6)$$

Eq. (6) can be used for the estimation of weights α and variances ψ. It can be equated to minimize the Negative Log Likehood $L_{min}(\alpha,\psi)$, transforming into the optimization problem:

$$L_{min}(\alpha,\psi) = \sum_{n=1}^{N} \left(\frac{|Z_q(n,k)|^2}{\psi_q(n,k)} + log\,\pi\psi_q(n,k)\right) \quad (7)$$

The model is continuously iterated to optimize the weights α and variance ψ to obtain the best results of the dereverberation model. After this process of model parameter optimization, the desired dereverberation signal $Z_q(n,k)$ can be calculated from Eq. (4).

III. Convolutional Time-Domain Speech Separation Network

In this paper, Conv-Tasnet is used for speech separation of the dereverberation signal.

A. Time Domain

Conv-Tasnet consists of three main components, the encoder, the separator and the decoder [13]. The mixture signal consists of C independent signals $s_1(t), s_2(t), \dots, s_C(t)$ superimposed in the time domain, which can be expressed as:

$$s(t) = \sum_{i=1}^{C} s_i(t) \quad (8)$$

In the time domain, our goal is to derive the separated $s_1(t), s_2(t), \dots, s_C(t)$, $C = 2$ in this paper.

B. Convolutional Encoder-Decoder Structure

The encoder which input mixture signal, and performs feature extraction and segmentation to map a segment of the signal to a higher dimension. The decoder performs the reverse operation on the masked signal, transforming the higher dimensional signal into the same dimension as the input signal and merging the segments.

In this network, given up the legacy method of speech separation is not used to estimate the time-frequency part of the signal by STFT. Instead, a convolutional encoder is introduced directly to extract the features of the speech. The mixture signal $s(t)$ can be divided into overlapping frames $\boldsymbol{\varepsilon_n}(t) \in R^{1\times L}$, where $n = 1,2,\dots,N$, and N represents the number of frames, "1" represents the number of channels of the input mixture signal, and L represents the length of the overlapping frames. With the 1-D convolution module, the overlapping frames $\boldsymbol{\varepsilon_n}(t)$ obtains a high-dimensional representation $\boldsymbol{w} \epsilon R^{1\times B}$. It is expressed by matrix multiplication as:

$$\boldsymbol{w} = \text{Relu}(\boldsymbol{\varepsilon_n v}) \quad (9)$$

where the dimension of the encoder matrix $\boldsymbol{v}$ is $B \times L$, B represents the basis function of the encoder. The Linear rectification function (Relu$(\cdot)$) as activator ensure that the output values are all positive.

The convolutional decoder is a 1-D transposed convolutional module that reconstructs the overlapping frames $\boldsymbol{\varepsilon_n}(t)$ from the system input $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{n}} \in R^{1\times L}$. $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{n}}$ represents the reconstruction of $\boldsymbol{\varepsilon_n}$, which can be expressed in the same way as a matrix multiplication:

$$\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{n}} = \boldsymbol{wu} \quad (10)$$

where the decoder matrix $\boldsymbol{u} \in R^{B\times L}$, B represents the basis function of the decoder. The overlapping parts of the reconstructed signals are superimposed to obtain the output of the model.

This convolutional encoder-decoder structure is used in this network to operate on overlapping frames by convolution and transposition convolution modules. It is beneficial to reduce latency and improve model accuracy.

C. Convolutional Separation Process

The separation part is the core of this network, and its basic idea is to estimate different masks by training the mixture signals.

The system multiplies the mask with the result of the convolutional encoder. It is further reconstructed by the convolutional decoder to get the separated signal estimated by the system, and this process is shown in Fig. 4.

The mask $\boldsymbol{m_i} \epsilon R^{1\times B}$, $i = 1,2,\dots,C$, C represents the number of sources. The mask $\boldsymbol{m_i}$ is an estimate of the C vectors for each frame of the mixture signal, subject to the following conditions:

$$\sum_{i=1}^{C} \boldsymbol{m_i} = 1_{1\times B} \quad (11)$$

where the elements are in $\boldsymbol{m_i}$ and take values from 0 to 1.

After the mask is obtained by the network, the weights of the mask are assigned to each source to acquire an initial estimation of the source $\boldsymbol{y}_i \in R^{1\times B}$, which is calculated as follows:

$$\boldsymbol{y}_i = \boldsymbol{w} \odot \boldsymbol{m_i} \tag{12}$$

where '$\odot$' represents the operation of multiplying the elements of two matrices at corresponding positions.

The separated source waveform which is an estimate of the ith initial source can be obtained as:

$$\hat{\boldsymbol{s}}_i = \boldsymbol{y_i u} \tag{13}$$

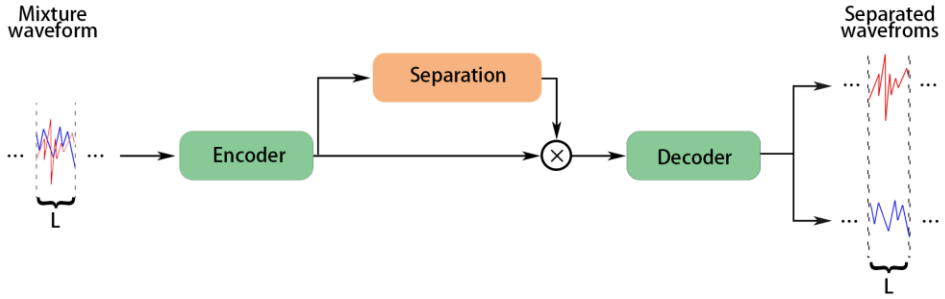

Figure 4. Conv-Tasnet block diagram.

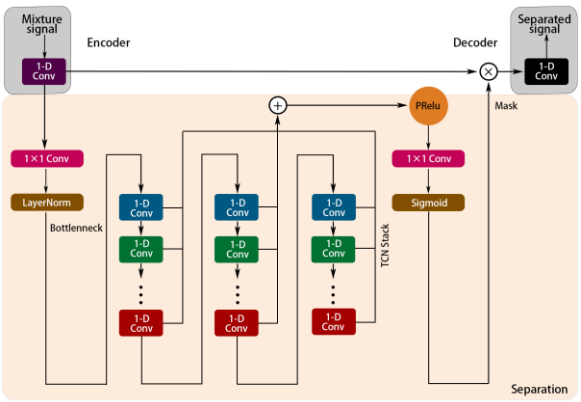

Figure 5. System flowchart of Conv-Tasnet.

Specifically, during the separation process, the feature dimension of $\boldsymbol{w} \epsilon R^{1\times \boldsymbol{B}}$ is reduced by a LayerNorm and a 1×1 Conv layer (Bottleneck layer). And then a three-layer Time-domain Convolutional Network (TCN) is used to increase the receptive field by stacking eight times 1-D convolutional block in each layer [14]. Thus, the long-time speech signals are modeled for higher classification accuracy. Fig. 5. Illustrates the system flowchart of Conv-Tasnet. In Fig. 5, the gray part represents the encoder and decoder, while the orange part represents the separation module. In the TCN, each 1-D convolutional block with different colors represents exponential growth due to its different dilation factors to satisfy the long-term dependence of the speech signal on the network. Each of these convolutional blocks requires the same length of input and is filled with "0" if it is insufficient.

The Fig. 6 shows the structure of 1-D convolutional block. Multiple 1-D convolutional blocks are connected to ensure a sufficiently large receptive field to exploit the long-range correlation of the speech signal. The output of one block is used as the input of the next block. After the expansion of the feature dimension, Depthwise convolution (D-conv) is adopted to perform collisional convolution in the time dimension. Skip-connection and Output are obtained by activation function (parametric rectified linear unit, *PRelu*) and normalization operation. In order to improve the fitting capacity of the model, $PRelu(\cdot)$ is expressed as follows:

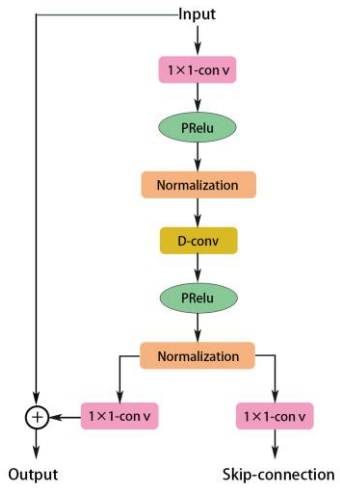

Figure 6. 1-D convolutional block.

$$PRelu(X) = \max(0, X) + a \times \min(0, X) \tag{14}$$

where $\max(0, X)$ represents the slope of 1 when the input is positive, and $a \times \min(0, X)$ represents the slope of a (learnable parameter) when the input is negative. All the Outputs passed to the *PRelu*, 1×1 Conv and sigmoid for mask $\boldsymbol{m_i}$ estimation. The estimated clean sources are obtained by reconstructing the hybrid encoder.

D. Loss Function

In this paper, the scale-invariant source-to-noise ratio (SI-SNR) is chosen as the loss function. It does not misestimate the similarity between the estimated clean source and the original clean source [15]. This is accomplished by projecting the estimated clean source to the vertical direction of the true vector (i.e., $\boldsymbol{x}$) and

suppressing the undesirable effects due to signal variations by regularization, as follows:

$$\boldsymbol{x} = \frac{\langle \hat{s}, s \rangle}{\|s\|^2} \boldsymbol{s} \tag{15}$$

where $\hat{\boldsymbol{s}} \in R^{1\times L}$ is the estimated clean source, $\boldsymbol{s} \in R^{1\times L}$ is the original clean source, $\langle \hat{\boldsymbol{s}}, \boldsymbol{s} \rangle$ is inner product operation. To ensure scale invariance, both $\hat{\boldsymbol{s}}$ and $\boldsymbol{s}$, required to be normalized (zero-mean normalization).

And then the SI-SNR, which is the evaluation metric of the speech separation system, is calculated as follows:

$$\mathrm{r}_{\mathrm{SI-SNR}} = 10\, log_{10} \frac{\|\boldsymbol{x}\|^2}{\|\hat{\boldsymbol{s}} - \boldsymbol{x}\|} \tag{16}$$

A larger value of $\mathrm{r}_{\mathrm{SI-SNR}}$ means a better separation performance.

IV. Experimental Procedure and Analysis of the Result

A. Experiment

In this paper, the Wall Street Journal dataset of two speakers (WSJ0-2mix) is selected. The dataset consists of a training set of about 30 hours of mixture signals, a validation set of about 10 hours and a test set of about 5 hours. Two speakers were randomly selected from a Wall Street Journal dataset (WSJ0) containing multiple speakers. The validation sets and test sets were generated with two different speakers at a Signal-to-Noise ratio (SNR) from -5dB to 5dB. Special processing was also done on the data of the test set. The First-order Ambinsonics (FOA) microphone was used for recording, four channels recorded signals from FOA microphone were simulated with different reverberation times under a two-source environment [16]. Different reverberation times (T60) of 0 ms, 150 ms, 300 ms, 450 ms, and 600 ms were simulated using ROOMSIM simulation software [17], and the information of the four channels front left up, front right down, back left down, and back right up was recorded in a room size of 6 m × 4 m × 3 m with the speech signal in the test set as the source. The signal was sampled at 8 kHz.

TABLE I. Network Setting

Symbol	Parameter	Description
N	512	Number of filters
L	16	Convolution kernel size
B	128	Number of 1 × 1 conv module channels
H	512	Number of channels in convolution blocks
P	3	The convolution kernel size of 1-D convolution module
X	8	Number of repetitions in each group
R	3	Number of repeats
Norm	Gln	The normalization method
Num-spks	2	Number of speakers
Activate	Relu	Activation function

In these experiments, a Conv-Tasnet is first trained to achieve the separation of pure speech based on the selected training and validation sets. Adam is selected as the optimizer; the learning rate is set to 0.001. Mixture signals of about 6s from two speakers were trained for 100 epochs. The settings of this network hyperparameters are shown in Table I.

After 100 epochs, the network model with the best results is saved. The four-channel recorded signal simulated by ROOMSIM is passed through the WPE method (relevant parameters are listed in Table II).

A four-channel output is generated, and the summation of the signals from each channel is extracted and fed into the Conv-Tasnet trained model for further testing (the sum signal is equivalent to an omnidirectional signal, containing all source information in the scene).

TABLE II. Parameters of WPE Method

Symbol	Parameter	Description
Channels	4	Number of input channels
Out-num	4	Number of output channels
P	15	Number of prediction order
L_P	2	Number of prediction step
Frame-size	2	Length of the frame
Overlap	0.5	The overlap factor between adjacent frames

B. Result Analysis

In this paper, the proposed method and the existing Conv-Tasnet method [13] were evaluated using Multi-Stimulus Test with Hidden Reference (MUSHRA) and Short-Time Objective Intelligibility (STOI), respectively [18, 19]. In order to compare the impact of reverberation on the speech separation system, two types of speech to be evaluated were selected. One is the separation result obtained by directly feeding the mixture signal into the speech separation system (Conv-Tasnet), and the other is the separation result obtained by the proposed method (i.e., removing the reverberation of the mixture signal first and then using a Conv-Tasnet to separate the mixture signal).

The MUSHRA listening test is a subjective evaluation method, in which the original signal is set as the reference (upper limit) and the signal that cannot be separated properly is set as the anchor factor (lower limit). The range of MUSHRA is 0–100, and the higher the value, the better the perceived quality is. In this paper, 15 listeners were selected for double-blind listening to evaluate the separation performance under different reverberation times, and all the measured data were averaged as shown in Fig. 7.

In addition, STOI is used as an objective evaluation method to score the speech to be evaluated by comparing the target speech with the speech to be evaluated. The value of STOI ranges from 0 to 1, and the higher the value, the better the intelligibility of the separation speech is. The separation performance of the proposed method was

evaluated under different reverberation times, and the average test results were shown as shown in Fig. 8.

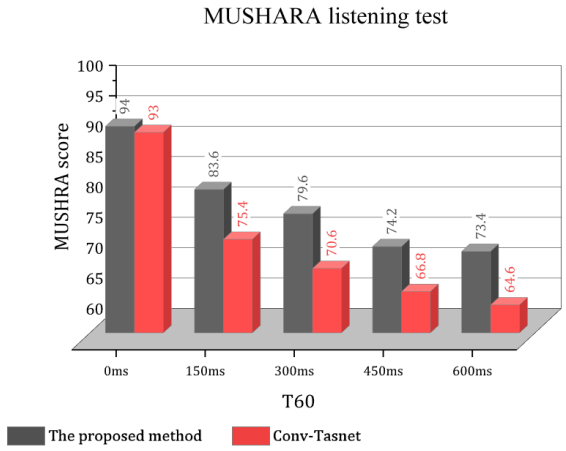

Figure 7. Results of MUSHRA listening test with 95% confidence intervals.

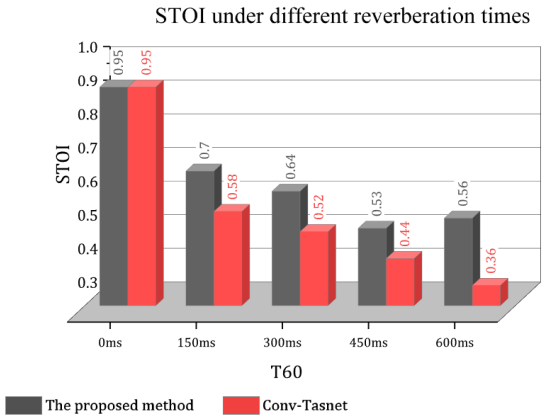

Figure 8. Results of STOI.

From Fig. 7 and Fig. 8, it is easily to figure out that the MUSHRA score as well as the STOI values keep decreasing as the reverberation time increases. The proposed method scored 8.4% higher in the MUSHRA and the 13.3% higher in the STOI test compared to the existing Conv-Tasnet method [13]. At the same time, the rate of decline becomes slower and more robust. The results show that this paper obtains better speech separation results that are more suitable for complex acoustic conditions. It shows that the proposed method is effective.

V. Conclusion

In this paper, a multi-speaker speech separation methods combining WPE method and Conv-Tasnet is proposed. First, the WPE method achieves reverberation removal by predicting the late reverberant component of the signal. Then, Conv-Tasnet is used to model directly in the time domain, and speech separation is achieved by the mask learned by TCN. The experimental results show that the proposed method can achieve good performance under different reverberation conditions. At the same time, no priori knowledge of the sound field environment is required. The generalization performance of the model is to be tested on other data sets.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Chunxi Wang performed the whole research and wrote the paper; Maoshen Jia provided support to the writing and experiments; Yanyan Zhang analyzed the data; Lu Li performed some experiments. The authors read and approved the final version of the paper.

Funding

This work was supported by the National Natural Science Foundation of China under Grants (61971015) and Beijing Natural Science Foundation (No. L223033).

References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

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

[3] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 46–50.

[4] X. Hu, K. Li, W. Zhang, *et al.*, "Speech separation using an asynchronous fully recurrent convolutional neural network," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22509–22522, 2021.

[5] M. Gao, Y. Gao, and F. Pei, "DNN-based speech separation with joint improved distortion constraints," in *Proc. 2021 14th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, China, 2021, pp. 5–8.

[6] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sept. 2010.

[7] C. J. Steinmetz, V. K. Ithapu, and P. Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," in *Proc. 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2021, pp. 221–225

[8] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[9] G. Li, S. Liang, S. Nie, and W. Liu, "Adaptive dereverberation using multi-channel linear prediction with deficient length filter," in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 556–560.

[10] D. Liang, M. D. Hoffman, and G. J. Mysore, "Speech dereverberation using a learned speech model," in *Proc. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Limassol, Cyprus, 2015, pp. 1871–1875.

[11] K. A. Karawi and D. Y. Mohammed, "Early reflection detection using autocorrelation to improve robustness of speaker verification in reverberant conditions," *International Journal of Speech Technology*, vol. 22, no. 4, pp. 1077–1084, 2019.

[12] T. Nakatani, B. H. Juang, T. Yoshioka, *et al.*, "Speech dereverberation based on maximum-likelihood estimation with

time-varying Gaussian source model," *IEEE Transactions on audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1512–1527, 2008.
[13] 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.
[14] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 6875–6879.
[15] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *Proc. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Singapore, 2019, pp. 260–267.
[16] L. Li, M. Jia, and J. Wang, "DOA estimation of multiple speech sources based on the single-source point detection using an FOA microphone," *Applied Acoustics*, vol. 195, pp. 1–16, 2022.
[17] D. R. Campbell, K. J. Palomaki, and G. J. Brown, "A MATLAB simulation of 'shoebox' room acoustics for use in research and teaching," *Computing and Information Systems*, vol. 9, no. 3, pp. 48–51, 2005.
[18] M. Schoeffler, F. R. Stöter, B. Edler, and J. Herre, "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the ITU-R recommendation BS. 1534 (MUSHRA)," in *Proc. 1st Web Audio Conference*, 2015, pp. 1–6.
[19] H. Zhang, X. Zhang, and G. Gao, "Training supervised speech separation system to improve STOI and PESQ directly," in *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Canada, 2018, pp. 5374–5378.