EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY

Open Access

Lightweight target speaker separation network based on joint training

Jing Wang[1], Hanyue Liu[1], Liang Xu[1], Wenjing Yang[1], Weiming Yi[2*] and Fang Liu[2]

Abstract

Target speaker separation aims to separate the speech components of the target speaker from mixed speech and remove extraneous components such as noise. In recent years, deep learning-based speech separation methods have made significant breakthroughs and have gradually become mainstream. However, these existing methods generally face problems with system latency and performance upper limits due to the large model size. To solve these problems, this paper proposes improvements in the network structure and training methods to enhance the model's performance. A lightweight target speaker separation network based on long-short-term memory (LSTM) is proposed, which can reduce the model size and computational delay while maintaining the separation performance. Based on this, a target speaker separation method based on joint training is proposed to achieve the overall training and optimization of the target speaker separation system. Joint loss functions based on speaker registration and speaker separation are proposed for joint training of the network to further improve the system's performance. The experimental results show that the lightweight target speaker separation network proposed in this paper has better performance while being lightweight, and joint training of the target speaker separation network with our proposed loss function can further improve the separation performance of the original model.

Keywords Target speaker separation, Lightweight network, Loss function, Joint training

1 Introduction

Speech signal processing techniques have a wide and meaningful application in daily life. In recent years, research on speech signal processing techniques has garnered more attention as the demand for remote collaboration and telecommuting has increased. The human auditory system can distinguish between different sounds, and can discern speech of interest from speech of disinterest even in noisy environments. In practical applications of human-computer interaction, the useful speaker's voice is often not pure, resulting in unsatisfactory practical applications for users. For example, some back-end speech signal processing systems, such as voice wake-up systems, rely on the front-end system to remove interfering voices and noises to perform better preprocessing on speech signals. Therefore, it is hoped that machines can also distinguish different sounds like humans to improve performance in noisy environments. This makes the need for speech separation technology more urgent.

The purpose of speech separation is to remove distracting human speech and background noise from mixed speech, which was introduced by Cherry's "cocktail party problem" in 1953 [1]. In real applications, the interference sound components in mixed speech are complex, and interference such as noise is often non-stationary. Meanwhile, the development of hardware technology, such as chips, enables the realization of speech separation technology with large computational load. Therefore, speech separation technology based

*Correspondence:
Weiming Yi
ywm@bit.edu.cn
[1] School of Information and Electronics, Beijing Institute of Technology, Beijing, China
[2] Key Laboratory of Language, Cognition and Computation Ministry of Industry and Information Technology, School of Foreign Languages, Beijing Institute of Technology, Beijing, China

on deep learning has rapidly developed in recent years [2–11] and achieved significant performance improvements compared with traditional speech separation technology [12–14]. However, since not all speakers' voices in mixed speech are necessary in some cases, separating all speakers' voices in mixed speech may lead to an unnecessary waste of computing resources and increase the system's delay, especially when only one target speaker's voice is needed. Target speaker separation is proposed to solve these problems. By confirming the target speaker that needs to be separated in advance and adding the voiceprint feature of the target speaker in the separation process for guidance, the targeted separation of specific speech components is achieved.

This paper mainly studies the frequency domain speech separation technology of single-channel speech signals based on deep learning. On the premise that the voiceprint information of the target speaker is obtained through steps such as speaker registration, the target speaker separation technology can achieve the following purposes: (1) with the assistance of the voiceprint information, the specific target speaker's voice components in the mixed speech can be separated to obtain high-quality voice signals of the target speaker; (2) continuous tracking and separation can be performed to realize the personalized memory of the device, which simplifies the steps by eliminating the need for subsequent screening and other steps; (3) the determined separation direction can reduce the workload of the speech separation system, thereby reducing the model's parameters so that the system can run efficiently on devices with limited computing resources and low system latency.

The deep learning-based target speaker separation method requires the target speaker's voiceprint information as a prerequisite for speech separation, and the performance of the model is influenced by the accuracy of the input voiceprint information. Therefore, the deep learning-based target speaker separation method is a combination of deep learning-based speaker recognition techniques and speech separation techniques. In order to further improve the separation performance of the system and put it into practical application, the existing deep learning-based target speaker separation methods still have the following problems that need to be solved: (1) the separation network based on recurrent neural network (RNN) has problems of large model, large number of parameters and long computational delay, which makes it difficult for the model to be implemented on small computing devices; (2) how to cooperate the training and optimization of the two sub-networks to realize performance improvement of the whole target speaker separation system.

In order to solve the above problems faced by target speaker separation, this paper makes improvements from the aspects of neural network structure and model lightweight, and improves the training method and loss function to improve the separation performance of the model at the same time. The innovations of this work mainly include the following two aspects:

1 A lightweight target speaker separation network based on long-short-term memory (LSTM) is proposed. To address the limitations of the existing LSTM-based target speaker separation network in terms of model size and actual computation volume, this paper introduces the idea of hierarchical level multiple extractions into the speaker separation network, increasing the intra-block and inter-block connections of neural network modules, and proposes a lightweight LSTM-based target speaker separation network. It can obtain better performance when the number of network parameters is reduced to about one-fifth of the baseline model, and can still maintain good performance when the number of network parameters is less than 1M.
2 A target speaker separation method based on joint training is proposed, which addresses the performance ceiling issue of the current single-channel frequency-domain method and the serious computational resource consumption of retraining when adapting to different scenarios. The two subsystems of the target speaker separation system, speaker registration and speaker separation system, are integrated and jointly perform parameter optimization during the training process. The joint training system can improve the separation performance of the whole system compared with the baseline model. At the same time, we propose a joint loss function based on speaker registration and speaker separation to be applied to the joint training system to further improve the performance of the joint training system.

The remainder of this paper is organized as follows. Section 2 discusses some related works and highlights the innovations of our method. Section 3 details the architecture of the lightweight target speaker separation network based on joint training. Section 4 introduces the experimental details. Section 5 discusses the experimental results of our method. Section 6 concludes this paper and proposes future research directions.

2 Related works

In recent years, deep learning-based techniques have gradually been applied in more and more separation tasks [15]. In 2013, Wang et al. [2] first proposed to use deep neural networks (DNN) [3] for feature extraction of speech signals in the field of speech separation, and then used a support vector machine (SVM) [4] as a classifier to predict the ideal binary mask (IBM) [5] of the target speech. The relevant experimental results show that this method can improve the effect of speech separation compared with the traditional shallow neural network. Since then, more and more researchers have begun to study and expand deep learning methods in the field of speech separation. In 2015, Huang [6] applied deep recurrent neural networks (DRNN) to the field of speech separation to further improve the performance of neural networks. In 2017, Chen and Wang et al. [7] used LSTM instead of DNN as a network for speech separation, and the RNN that considers the continuity of speech information in the time dimension can learn the input features more effectively. The experimental results show that LSTM can significantly improve the generalization performance of the model when the number of speakers increases.

In contrast to speech separation, target speaker separation refers to extracting the speech of the target speaker from background or mixed speech. In order to achieve this goal, one way is to apply the speech separation method to mixed speech to separate the speech of different speakers first, then extract the target speech. However, this method is resource-consuming and it is difficult to decide which speech to extract. A more common method is to treat it as a binary classification task, where the positive class is the speech of the target speaker, and the negative class is the combination of speech of all the interfering speakers and noises. In 2017, Zmolikova et al. [16] used neural networks to estimate masks of the target speakers and used these masks to derive beamformer filters, which are used to extract the target speech. By informing the neural network about the target speaker in advance to make its layer dependent on the speaker's characteristics, the separation of the target speaker can be achieved. Later, Zmolikova et al. [17] proposed a joint scheme using a sequence summarizing system, which combines the learning of the speaker representation with the learning of speaker separation and shows better performance than extracting the speaker representation separately. In 2018, Wang et al. [18] proposed a novel "deep extractor network" to extract the features of both the target speech and the mixed speech, which shows good feature extraction performance and can effectively recover target speech from the mixed speech even given a very short utterance. In 2019, Wang et al. [19] proposed to train the speaker recognition network and spectrogram masking network separately to separate the speech of the target speaker from multi-speaker signals, which significantly reduces the speech recognition word error rate (WER) on multi-speaker signals. In 2020, Xu et al. [20] proposed a time-domain speaker extraction network (SpEx) that converts the mixture speech into multi-scale embedding coefficients instead of decomposing the speech signal into magnitude and phase spectra, avoiding phase estimation. The target speaker's speech is reconstructed from the masked embedding coefficients by speech decoder. In 2022, He et al. [21] proposed an improved target speaker extractor that combines the time domain and frequency domain. By leveraging the time-domain modeling ability of WaveUNet and the frequency-domain modeling ability of convolutional recurrent network (CRN), the target speaker can be tracked well.

This paper focuses on frequency domain single channel target speaker separation technique based on deep learning, which mainly includes two subsystems: the speaker registration system and the speaker separation system. The target speaker's voiceprint feature information needs to be obtained in advance by the speaker registration system. In the speaker separation system, by using the target speaker's voiceprint information and the features of mixed speech signals as the input of the network, the desired clean speech of the target person is finally output. In most DNN-based target speaker separation systems, the parameters of the speaker registration subsystem are typically held constant during both training and testing of the speaker separation system. These parameters are pretrained and embedded in the target speaker separation system in advance. This approach limits system performance and hampers adaptability to changes in application scenarios. Additionally, many small terminal devices such as cell phones and stereos lack sufficient computational resources to allocate to target speaker separation systems when compared to computers. To address these problems, this paper proposes a target speaker separation method based on lightweight LSTM and joint training, which greatly reduces the parameters and computation amount of the speaker separation network. At the same time, the speaker registration subsystem is added to the learning of the current problem, and two joint loss functions based on speaker registration and speaker separation are proposed for this system to further improve the performance of the target speaker separation system.

3 Proposed network

The overall structure of the target speaker separation network proposed in this paper is shown in Fig. 1, which consists of a speaker registration system and a

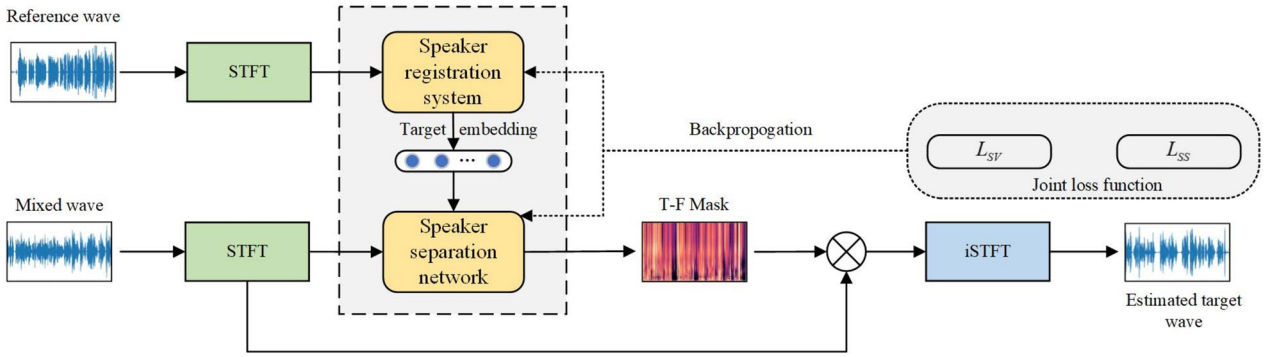

Fig. 1 The architecture of our proposed system

lightweight speaker separation network, and the system is equipped with corresponding loss function.

3.1 Speaker registration system

The purpose of the speaker registration network is to extract the target embedding from the original reference audio. In this paper, we refer to the speaker recognition network used by Chung et al. [22] in training the speaker registration network [23]. ResnetSE-34, a 34-layer neural network based on Resnet [24] combined with channel attention, is used in our speaker registration system. The dimension of the output embedding is 256 dimensions. The use of Resnet enables the deep neural network to maintain good learning performance while effectively avoiding the gradient disappearance or explosion during the training process. Traditional feature extraction models often perform statistical integration after extracting multiple feature components, such as in the extraction process of the embedding d-vector [25], where an average operation is performed on the feature vectors proposed in each frame to obtain the optimal embedding. In order to optimize the feature extraction and learning of the adopted basic network framework Resnet, and improve the efficiency and effectiveness of feature extraction, this paper introduces a channel attention mechanism to Resnet. Channel attention mechanism obtains the weight information of each channel feature through learning so that the weights of different feature vectors of multiple channels can be considered comprehensively, instead of simply giving equal weights to find the average. The channel attention network used in this paper is the squeeze-and-excitation (SE) block [26], which was originally proposed in the field of image processing, and can be embedded and used in many different types of networks.

3.2 Speaker separation system

In this paper, we propose a lightweight LSTM-based target speaker separation network (hereinafter referred to as "lightweight speaker separation network"), which is composed of lightweight LSTM-based speaker separation modules (hereinafter referred to as "lightweight speaker separation module"). The proposed lightweight speaker separation network consists of n lightweight speaker separation modules, an FC input layer, and an FC output layer. The network structure is shown in Fig. 2.

The input of the network is the target speaker's voiceprint embedding and the amplitude spectrum of the mixed speech, and the output is the T-F Mask of the clean speech of the target speaker. The lightweight network structure proposed in this paper employs LSTM as the main structure, which is particularly adept at handling time series information. The two inputs are not directly combined. Besides, the voiceprint embedding of the target speaker is added to the speaker separation modules at all levels from front to back multiple times to achieve multiple extractions from the target speaker's speech components in the mixed speech amplitude spectrum. Through the design of multi-level extractions within and between modules and skip connections within the block, the proposed lightweight speaker separation network can reduce the amount of network parameters while maintaining the learning performance of the network, achieving the purpose of network lightweight.

The structure of the lightweight speaker separation module is shown in Fig. 3, which consists of fully connected (FC) layers, LSTM, and a conditional layer normalization (CLN) [27]. The FC layer is mainly responsible for dimensional transformation of the input and the transmitted information in the network. LSTM is the core component of the module and is responsible for effective learning of the input feature information. CLN is a variant of the conditional batch normalization

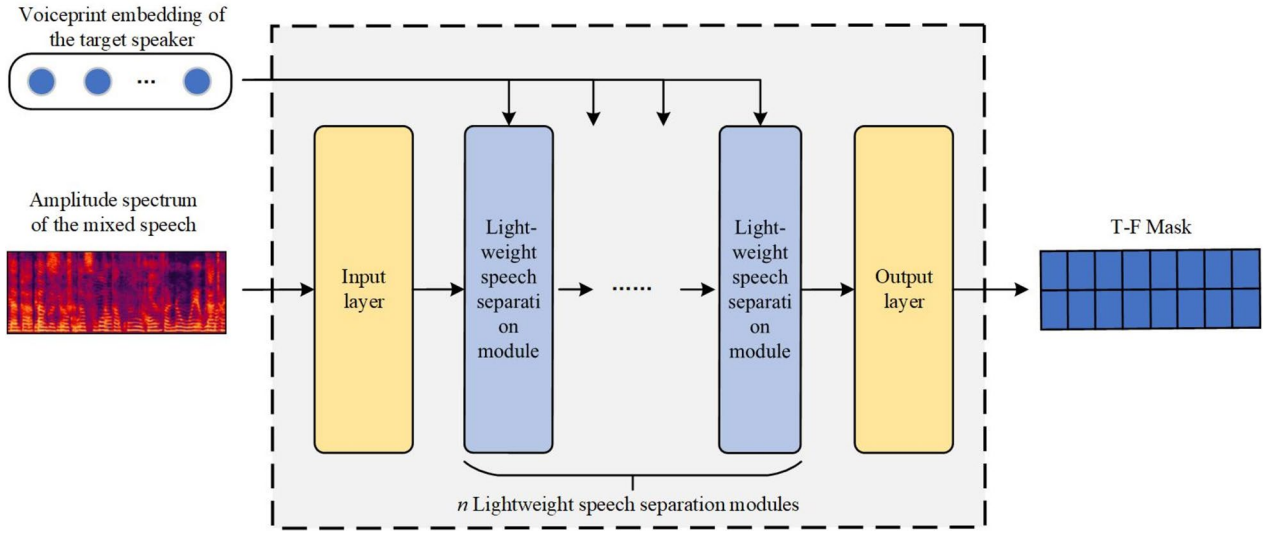

Fig. 2 Structure of the lightweight speaker separation network based on LSTM

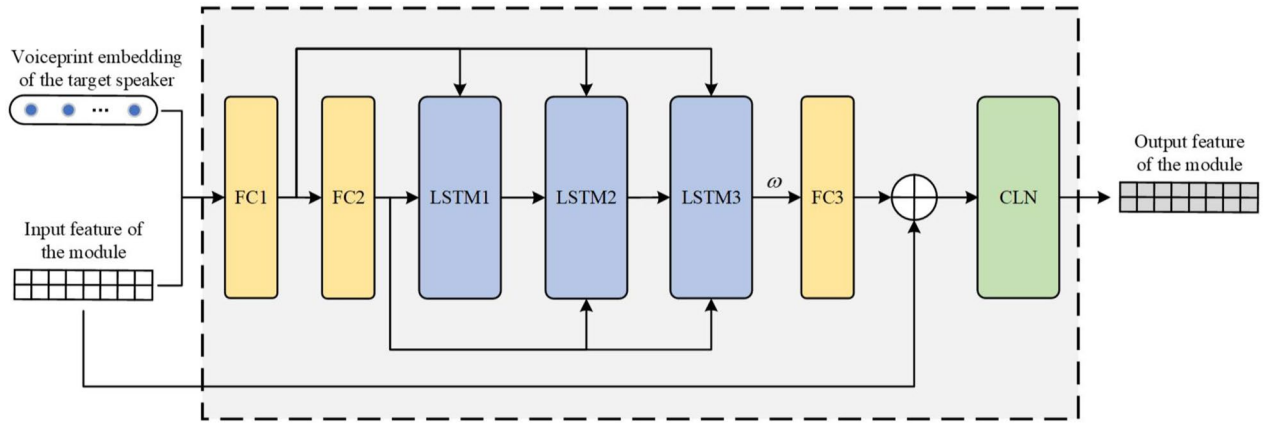

Fig. 3 Structure of the lightweight speaker separation module based on LSTM

(CBN). Compared with the general layer normalization (LN) [28], CLN can make the parameters of the network learnable, allowing the input information to be personalized and standardized. This enables the output results to be more closely aligned with the target direction.

In the lightweight speaker separation module, the input of the module is the voiceprint embedding of the target speaker and speech input features, and the output is the speech features calculated and processed by the module. The neural network module consists of 7 layers of neural networks, including 3 layers of FC, 3 layers of LSTM, and one layer of CLN. In the intra-module network, the input information is processed by FC1 and then passed to FC2 and 3-layer LSTM at the same time, and the transmitted information is processed by FC2 and then passed to the 3-layer LSTM at the same time. Effective information is extracted using hierarchical multiple extractions, which improves the effectiveness of the output information ω of the last layer of LSTM, thereby achieving good feature learning performance while reducing the number of neurons in each layer of the neural network, and realizing lightweight of the speaker separation module. Finally, the output of FC3 and the original input features of the current module are added and input to the CLN to obtain the final module output features after conditional normalization.

According to the number of modules used and the size of the modules, this paper proposes two kinds of lightweight speaker separation networks: Light Target Speaker Separation Net (LTSS Net) and Super-Light Target Speaker Separation Net (S-LTSS Net). They are embedded into the target speaker separation system to form two lightweight systems for experiments. The details of the two lightweight speaker separation

Table 1 Parameters of lightweight speaker separation network

Network	Number of modules (*n*)	Parameters of LSTM (1/2/3)	Parameters of network
LTSS Net	16	48/96/192	4.64M
S-LTSS Net	4	24/48/96	0.55M

networks are shown in Table 1. The number of modules and other parameters are decided through experiments, which is the minimum parameter quantity while ensuring the performance of the model. The parameters of the network include all weight matrices, bias vectors and optimizer parameters in the network, which are computed by PyTorch.

3.3 Loss function

The joint loss function proposed in this paper is composed of speaker verification loss (SVL) and speaker separation loss (SSL). The two loss functions are derived from the speaker registration task and the speaker separation task respectively. This paper proposes two joint loss functions based on the two tasks.

The first joint loss function (Joint Loss-1, $L_{Joint-1}$) uses the classification-based loss function AM-Softmax Loss [29] as SVL, and the time-frequency (T-F) domain combinative loss function [30] ($L_{Combinative}$) composed of RMSE and Si-SNR as SSL. The calculation formula of $L_{Joint-1}$ proposed in this paper is as follows.

$$\begin{aligned} L_{Joint-1} &= \alpha L_{SV} + \beta L_{SS} \\ &= \alpha L_{AM-Softmax} + \beta L_{Combinative} \\ &= \alpha L_{AM-Soft\max} \\ &+ \beta(\omega_1 L_{RMSE} + \omega_2 L_{Si-SNR}) \end{aligned} \tag{1}$$

where α and β are the weights of SVL and SSL, respectively. ω_1 and ω_2 are the weights of the frequency domain loss function and the time domain loss function in the integrated loss function $L_{Combinative}$, respectively. After several performance test experiments, we set $\alpha = 0.2$, $\beta = 0.8$, $\omega_1 = 0.5$, and $\omega_2 = 0.5$.

The second joint loss function (Joint Loss-2, $L_{Joint-2}$) comprehensively uses the classification-based loss function LMCL [31] and the vector-based loss function triplet loss [32] by assigning weights as the loss function of speaker verification. The T-F domain combinative loss function $L_{Combinative}$ is used as the loss function of speaker separation. The calculation formula of $L_{Joint-2}$ proposed in this paper is as follows.

$$\begin{aligned} L_{Joint-2} &= \alpha L_{SV} + \beta L_{SS} \\ &= \alpha L_{SV} + \beta L_{Combinative} \\ &= \alpha\left(\psi_1 L_{LMC} + \psi_2 L_{Triple}\right) \\ &+ \beta(\omega_1 L_{RMSE} + \omega_2 L_{Si-SNR}) \end{aligned} \tag{2}$$

After many performance test experiments, we set $\alpha = 0.2$, $\beta = 0.8$, $\psi_1 = 0.5$, $\psi_2 = 0.5$, $\omega_1 = 0.5$, and $\omega_2 = 0.5$.

4 Experiments

4.1 Datasets

In this paper, when training the speaker registration model, multiple open-source datasets are used to form the training dataset, including VoxCeleb1 [33], VoxCeleb2 [34], CNCeleb [35], LibriSpeech [36], Aishell2 [37], and ST-CMDS, with a total of 10,920 speakers. In the testing phase, the test set from the VoxCeleb1 [33] dataset is used as the test data, which includes 40 speakers. All the speech data are pre-normalized when training the speaker registration model. Noise or reverberation randomly selected from the Musan dataset [38] is added during the training process, and the signal-to-noise ratio (SNR) when performing mixing is randomly selected from the set {0 dB, 5 dB, 10 dB}.

The open-source dataset LibriSpeech [36] is used in the target speaker separation experiments, and the train set and the test set include 2338 speakers and 73 speakers, respectively. The mixed speech is generated in real-time in two steps during the model training and performance testing: (1) one speaker is randomly selected from the dataset as a target speaker, and one registered speech and one training speech are selected from the speaker's speech data for speech registration and separation training, respectively; (2) another speaker is randomly selected from the dataset as an interfering speaker, and one speech is selected from the speaker's speech data to be mixed with the target speaker's training speech. The signal-to-intensity ratio (SIR) is randomly selected from the set {−5 dB, 0 dB, 5 dB, 10 dB}.

In practice, all data are cut and linked to ensure that the length of each data is not less than 10 s, and the utterances used for registration and separation are 10 s and 3 s, respectively.

4.2 Implementation details

In this paper, the sampling rate of all speech data is 16 kHz. A 512-point FFT is performed, and the frame length and frame shift are set to 400 and 160, respectively. Adam optimizer is used, and the initial learning rate is set to 0.001. All neural network models are trained for 150 epochs, and the experimental results are the results after the models converge.

4.3 Evaluation metrics

The evaluation metrics of the target speaker separation task are mainly divided into two types: subjective and objective, and objective evaluation metrics are mainly used in this paper. The experimental results are evaluated using several objective evaluation metrics commonly used in the field of speech separation and speech enhancement. These metrics include segment source-to-noise ratio (SSNR), perceptual evaluation of speech quality (PESQ) [39], short-time objective intelligibility (STOI) [40], log spectral distance (LSD) [41], and three comprehensive mean opinion score (MOS) [42] metrics for speech quality: MOS predictor of intrusiveness of background noise (CBAK), MOS predictor of speech distortion (CSIG), MOS predictor of overall processed speech quality (COVL). Among these metrics, except for LSD, higher values represent better performance. Since the loss function used in the neural network training experiments includes Si-SNR, it is not used as the evaluation metric of the experiments in this paper.

5 Results and discussion

5.1 LSTM-based lightweight target speaker separation

The experiments in this section focus on demonstrating and analyzing the separation performance of proposed target speaker separation networks of different sizes. ResnetSE-34 trained with the generalized end-to-end (GE2E) [43] loss function is used as the speaker registration model. The speaker registration system is first trained with a larger dataset, after which the trained model is embedded and used in the target speaker separation. The baseline model for speaker separation is Voice-Filter [19]. We first reproduce the baseline model and then train it with the combinative loss function in the T-F domain. In the comparison experiments, the two lightweight speaker separation networks, LTSS Net and S-LTSS Net, proposed in this paper are compared with the baseline model, with model size being the only variable. The performance of the lightweight target speaker separation model proposed in this paper is explored by comparing its performance on the LibriSpeech test set.

Table 2 shows the four objective evaluation metrics, namely SSNR, PESQ, STOI, and LSD, for the separated speech. On the other hand, Fig. 4 shows the three comprehensive speech quality evaluation metrics in terms of MOS, namely CBAK, CSIG, and COVL. The first row in each table and the first group in each figure show the initial metric values obtained before the voice data is processed by the system.

From the results of speech metrics shown in Table 2 and Fig. 4, it can be concluded that the Voice-Filter model with combinative loss achieves high results in four metrics SSNR, PESQ, STOI, and LSD, and shows good performance in three MOS evaluation metrics. The seven metrics show that the baseline model can greatly exceed the application standards, but the large number of model parameters 9.86M makes it difficult to be applied on small devices with limited computational resources. The proposed LTSS Net model with combinative loss can achieve better performance with the number of parameters being reduced to 1/2 of the baseline model. The number of parameters of LTSS Net is only 4.64M, which meets the use conditions of some small

Table 2 Experimental results on Librispeech test set

Network	Parameters	SSNR	PESQ	STOI	LSD
None	–	1.61	1.26	0.64	7.16
Voice-Filter	9.86M	12.17	2.44	0.84	4.40
LTSS Net	4.64M	**12.33**	**2.48**	**0.86**	**4.39**
S-LTSS Net	**0.55M**	9.65	1.89	0.78	5.26

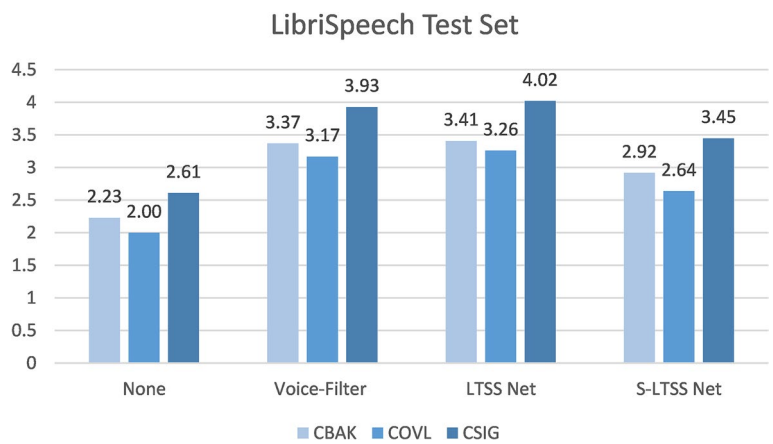

Fig. 4 Experimental results on Librispeech test set

devices with limited computing resources. The S-LTSS Net model is a super lightweight model proposed in this paper, which aims to reduce the number of model parameters to the limit while maintaining good system performance. As a super lightweight speaker separation model with only 0.55M parameters, the number of parameters is reduced by a factor of 18 compared with the baseline model Voice-Filter. Although the performance decreases compared with the baseline model, numerically all the evaluation metrics are greatly improved compared with the initial value, and the performance of target speaker separation is still good. As a super lightweight model with the number of parameters less than 1M, it is almost not limited by the computational resources of the device and can meet the conditions of many small devices.

5.2 Lightweight target speaker separation based on joint training

In this part of the experiment, ResnetSE-34 network model trained with GE2E loss function is used in the speaker registration system, and the super-lightweight model S-LTSS Net is used in the speaker separation system. The joint training-based target speaker separation is experimented based on them. During the training process, the pre-trained speaker registration subsystem no longer has fixed parameters, but is further trained with the whole system. The comparative experiments on the jointly trained system are conducted using the super lightweight model S-LTSS Net proposed in this paper, which is trained with the combinative loss and two joint loss functions (Joint Loss-1 and Joint Loss-2) proposed in this paper. The experimental details are shown in Table 3.

In this part of the experiment, the S-LTSS Net using combinative loss without joint training method is used as the baseline model. Four metrics, SSNR, PESQ, STOI, and LSD, of the separated speech are shown in Table 4. The three speech MOS evaluation metrics, CBAK, COVL, and CSIG, are demonstrated by the bar chart shown in Fig. 5.

The experimental results in Table 4 and Fig. 5 show that S-LTSS Net-2 based on joint training achieves an overall performance improvement over the baseline model on the LibriSpeech test set, with all seven metrics improved. Compared with the pure joint training method in S-LTSS Net-2, the S-LTSS Net-3 based on the joint loss function Joint Loss-1 can further improve the performance on the LibriSpeech test set, with five of the seven speech metrics improved, one remaining flat, and one slightly decreased. The experimental results show that the joint loss function Joint Loss-1 can improve the target speaker separation performance of the joint training method on S-LTSS Net, but the improvement is small and the stability is not high enough. The S-LTSS Net-4 based on Joint Loss-2 achieves a comprehensive and stable performance improvement compared with S-LTSS Net-2 on the LibriSpeech test set, and all seven metrics are improved. The experimental results demonstrate that Joint Loss-2 can significantly and consistently enhance the performance of the jointly trained S-LTSS Net, and outperforms Joint Loss-1.

Table 3 Details of the joint training experiment of S-LTSS Net

Network	Loss function	Parameters of network	Joint training or not
S-LTSS Net-1	Combinative loss	0.55M	N
S-LTSS Net-2	Combinative loss	0.55M	Y
S-LTSS Net-3	Joint Loss-1	0.55M	Y
S-LTSS Net-4	Joint Loss-2	0.55M	Y

Table 4 Experimental results of S-LTSS Net

Joint training or not	Loss function	SSNR	PESQ	STOI	LSD
–	–	1.61	1.26	0.64	7.16
N	Combinative loss	9.65	1.89	0.78	5.26
Y	Combinative loss	9.83	1.97	0.79	5.00
Y	Joint Loss-1	10.06	1.93	0.79	4.98
Y	Joint Loss-2	**10.12**	**2.05**	**0.81**	**4.97**

6 Conclusion

In this paper, we propose a lightweight network-based target speaker separation method that uses joint training to improve separation performance. By reducing the number and size of LSTM modules and incorporating multi-level extractions and hopping within and between blocks, our proposed network achieves network lightweight while maintaining learning performance. By jointly training the speaker registration and speaker separation networks and proposing a joint loss function, we further improve the separation performance. Our experimental results show that the super-lightweight target speaker separation network, S-LTSS Net, with a model volume of less than 1M, can run efficiently on small devices with limited computational resources while achieving good separation performance. Our joint training method based on Joint Loss-2 further improves the separation effect of the model.

The method proposed in this paper only studies the single-channel target speaker separation task in the

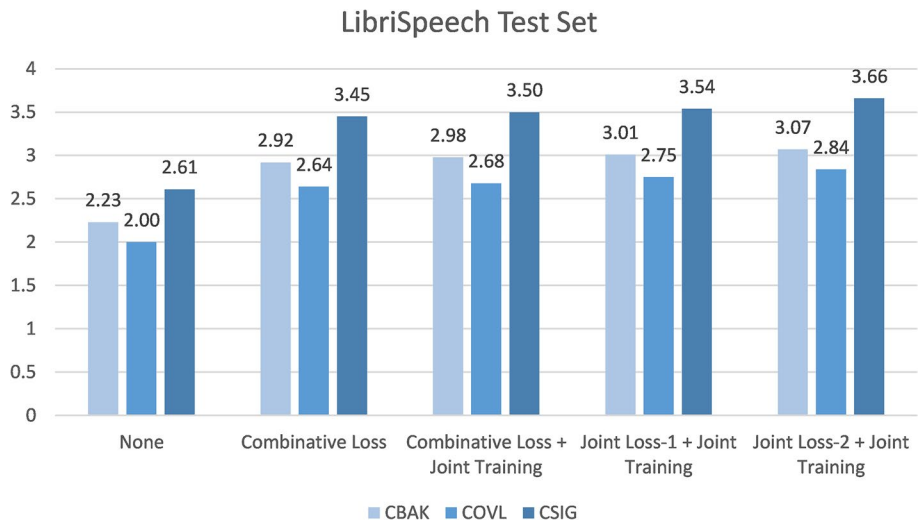

Fig. 5 Experimental results of S-LTSS Net

frequency domain in terms of joint loss function and lightweight network, and does not consider the dual-channel and time-domain methods. In the future, we will consider further experiments and improvements in areas such as dual-channel and time-domain speaker separation.

Abbreviations

LSTM	Long-short-term memory
RNN	Recurrent neural network
SVM	Support vector machine
IBM	Ideal binary mask
DRNN	Deep recurrent neural network
WER	Word error rate
CRN	Convolutional recurrent network
SE	Squeeze-and-excitation
FC	Fully connected
CLN	Conditional layer normalization
CBN	Conditional batch normalization
LN	Layer normalization
SVL	Speaker verification loss
SSL	Speaker separation loss
T-F	Time-frequency
SNR	Signal-to-noise ratio
SIR	Signal-to-intensity ratio
SSNR	Segment source-to-noise ratio
PESQ	Perceptual evaluation of speech quality
STOI	Short-time objective intelligibility
LSD	Log spectral distance
MOS	Mean opinion score
GE2E	Generalized end-to-end

Acknowledgements
We gratefully thank the reviewers and editors for their effort in the improvement of this work.

Authors' contributions
JW directed the project. JW conceived the algorithm. HL drafted the document. WY and LX wrote the software, executed the experiments. HL and JW completed the final manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by National Nature Science Foundation of China (Grant No.62071039) and Beijing Natural Science Foundation (Grant No.L223033).

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 14 June 2023 Accepted: 7 November 2023
Published online: 06 December 2023

References

1. E.C. Cherry, Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. **25**(5), 975–979 (1953)
2. D.L. Wang, J. Chen, Supervised speech separation based on deep learning: an overview. IEEE Trans. Audio Speech Lang. Process. **26**(10), 1702–1726 (2018)
3. A. Canziani, A. Paszke, E. Culurciello, An analysis of deep neural network models for practical applications (2016). http://arxiv.org/abs/1605.07678
4. W.S. Noble, What is a support vector machine? Nat. Biotechnol. **24**(12), 1565–1567 (2006)
5. D.L. Wang, On ideal binary mask as the computational goal of auditory scene analysis. In: Speech Separation by Humans and Machines, pp. 181–197. (Springer, Boston, 2005)
6. P.S. Huang, M. Kim, M. Hasegawa-Johnson, et al., Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE Trans. Audio Speech Lang. Process. **23**(12), 2136–2147 (2015)
7. J. Chen, D.L. Wang, Long short-term memory for speaker generalization in supervised speech separation. J. Acoust. Soc. Am. **141**(6), 4705–4714 (2017)
8. Y. Luo, N. Mesgarani, Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP). (IEEE, Calgary, 2018), p. 696–700

9. Y. Luo, N. Mesgarani, Conv-tasnet: Surpassing ideal time frequency magnitude masking for speech separation. IEEE Trans. Audio Speech Lang. Process. **27**(8), 1256–1266 (2019)
10. C. Lea, M.D. Flynn, R. Vidal, et al., Temporal convolutional networks for action segmentation and detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., (IEEE, Honolulu, 2017), p. 156–165
11. Y. Luo, Z. Chen, T. Yoshioka, Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP). (IEEE, Barcelona, 2020), p. 46–50
12. M.H. Radfar, R.M. Dansereau, A. Sayadiyan, Monaural speech segregation based on fusion of source-driven with model-driven techniques. Speech Commun. **49**(6), 464–476 (2007)
13. M.N. Schmidt, R.K. Olsson, Single-channel speech separation using sparse non-negative matrix factorization. In: Interspeech. (ISCA, Pittsburgh, 2006), p. 2–5
14. J.V. Stone, Independent Component Analysis: a Tutorial Introduction. (MIT Press, Cambridge, 2004)
15. J. Wang, H. Liu, H. Ying, C. Qiu, J. Li, M.S. Anwar, Attention-based neural network for end-to-end music separation. CAAI Trans. Intell. Technol. **8**, 355–363 (2023). https://doi.org/10.1049/cit2.12163
16. K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, T. Nakatani, Speaker-aware neural network based beamformer for speaker extraction in speech mixtures. In: Interspeech (2017)
17. K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, T. Nakatani, Learning speaker representation for neural network based multichannel speaker extraction. In: IEEE Autom. Speech Recognit. Understanding Workshop (ASRU). (IEEE, Okinawa, 2017), p. 8–15
18. J. Wang, J. Chen, D. Su, L. Chen, M. Yu, Y. Qian, D. Yu, Deep extractor network for target speaker recovery from single channel speech mixtures (2018). http://arxiv.org/abs/1807.08974
19. Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. Saurous, R. Weiss, Y. Jia, I. Moreno, Voicefilter: targeted voice separation by speaker-conditioned spectrogram masking, pp. 2728–2732 (2019). https://doi.org/10.21437/Interspeech.2019-1101
20. C. Xu, W. Rao, E.S. Chng, H. Li, Spex: Multi-scale time domain speaker extraction network. IEEE Trans. Audio Speech Lang. Process. **28**, 1370–1384 (2020)
21. S. He, H. Li, X. Zhang, Speakerfilter-pro: an improved target speaker extractor combines the time domain and frequency domain. In: 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 473–477 (2022). https://doi.org/10.1109/ISCSLP57327.2022.10037794
22. J.S. Chung, J. Huh, S. Mun, et al., In defence of metric learning for speaker recognition (2020). http://arxiv.org/abs/2003.11982
23. T. Kinnunen, H. Li, An overview of text-independent speaker recognition: from features to supervectors. Speech Commun. **52**(1), 12–40 (2010)
24. Y. Tai, J. Yang, X. Liu, Image super-resolution via deep recursive residual network. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (IEEE, Honolulu, 2017), p. 3147–3155
25. E. Variani, X. Lei, E. McDermott, et al., Deep neural networks for small footprint text-dependent speaker verification. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP). (IEEE, Florence, 2014), p. 4052–4056
26. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., (IEEE, Salt Lake City, 2018), p. 7132–7141
27. D. Lee, Z. Tian, L. Xue, et al., Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization (2021). http://arxiv.org/abs/2108.00449
28. J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization (2016). http://arxiv.org/abs/1607.06450
29. F. Wang, J. Cheng, W. Liu, et al., Additive margin softmax for face verification. IEEE Signal Process Lett. **25**(7), 926–930 (2018)
30. W.J. Yang, et al., A target speaker separation neural network with joint-training. In: 2021 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC). (APSIPA, Tokyo, 2021), p. 614–618
31. H. Wang, Y. Wang, Z. Zhou, et al., Cosface: large margin cosine loss for deep face recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., (IEEE, Salt Lake City, 2018), p. 5265–5274
32. A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification (2017). http://arxiv.org/abs/1703.07737
33. A. Nagrani, J.S. Chung, A. Zisserman, Voxceleb: a large-scale speaker identification dataset (2017). http://arxiv.org/abs/1706.08612
34. J.S. Chung, A. Nagrani, A. Zisserman, Voxceleb2: deep speaker recognition (2018). http://arxiv.org/abs/1806.05622
35. Y. Fan, J.W. Kang, L.T. Li, et al., Cn-celeb: a challenging Chinese speaker recognition dataset. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP). (IEEE, Barcelona, 2020), p. 7604–7608
36. V. Panayotov, G. Chen, D. Povey, et al., Librispeech: an ASR corpus based on public domain audio books. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP). (IEEE, Brisbane, 2015), p. 5206–5210
37. J. Du, X. Na, X. Liu, et al., Aishell-2: transforming mandarin ASR research into industrial scale (2018). http://arxiv.org/abs/1808.10583
38. D. Snyder, G. Chen, D. Povey, Musan: a music, speech, and noise corpus (2015). http://arxiv.org/abs/1510.08484
39. A.W. Rix, J.G. Beerends, M.P. Hollier, et al., Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP). (IEEE, Salt Lake City, 2001), p. 749–752
40. C.H. Taal, R.C. Hendriks, R. Heusdens, et al., A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP). (IEEE, Dallas, 2010), p. 4214–4217
41. A. Gray, J. Markel, Distance measures for speech processing. IEEE Trans. Audio Speech Lang. Process. **24**(5), 380–391 (1976)
42. R.C. Streijl, S. Winkler, D.S. Hands, Mean Opinion Score (MOS) revisited: methods and applications, limitations and alternatives. Multimedia Syst. **22**(2), 213–227 (2016)
43. L. Wan, Q. Wang, A. Papir, et al., Generalized end-to-end loss for speaker verification. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP). (IEEE, Calgary, 2018), p. 4879–4883

Publisher's Note