IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 May 2023, accepted 14 June 2023, date of publication 20 June 2023, date of current version 18 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3287868

RESEARCH ARTICLE

Speaker Verification Based on Single Channel Speech Separation

RONG JIN, MIJIT ABLIMIT, AND ASKAR HAMDULLA
School of Information Science and Engineering, Xinjiang University, Urumqi 830017, China
Xinjiang Key Laboratory of Signal Detection and Processing, Urumqi 830017, China

Corresponding author: Mijit Ablimit (mijit@xju.edu.cn)

This work was supported in part by the Strengthening Plan of National Defense Science and Technology Foundation of China under Grant 2021-JCJQ-JJ-0059, and in part by the Natural Science Foundation of China under Grant U2003207.

ABSTRACT In multi-speaker scenarios, speech processing tasks like speaker identification and speech recognition are susceptible to noise and overlapped voices. As the overlapped voices are a complicated mixture of signals, a target extraction method from this mixture is a good front-end solution for further processing like understanding and classifying. The quality of speech separation can be assessed by the noise ratio or subjective scoring and can also be assessed by accuracy of the downstream processing tasks like speaker identification. In order to make the separation model and speaker identification model more adapted to complex multi-speaker speech overlapping scenarios, this research investigates the speech separation model and incorporate with a voiceprint recognition task. This paper proposes a feature-scale single channel speech separation network connected to a back-end speaker verification network with MFCCT features, so the accuracy of speaker identification indicates the quality of speech separation task. The datasets are prepared by synthesizing Voxceleb1 data, and used for training and testing. The results show that using an objective downstream evaluation can effectively improve the overall performance, as the optimized speech separation model significantly reduced the error rate of speaker verification.

INDEX TERMS Speech separation, voiceprint recognition, speaker verification, multi-tasking.

I. INTRODUCTION

Various voices are overlapped in reality scenarios, and the target voice power-ratio (SNR) must be higher for humans to understand or identify the content. Voice separation techniques made it possible to separate the target contents from the mixture of voices even with low power ratios. The neural network models have shown to be highly effective [1], [2], [3], [4] on voice separation tasks. Lutati et al. [5] focused on the upper bound of single-channel speech separation and proposed SepIt networks, which showed good performance for multi-speaker speech separation with 5 and 10 speakers. Yang and Bao [6] combine RNN with CNN, and discussed separation performance and computational efficiency. Paturi et al. [7] researched long form reality conversational telephone speech, and proposed a speaker conditioned separator trained on speaker embeddings, in which the target speaker is extracted directly from the mixed signal using an over-clustering based approach. Thus, the overlapped voices are efficiently separated and used for several downstream applications such as ASR and speaker identification (SI). Usually, a target voice is extracted from a mixture of voices or noise in order to improve the intelligibility or accuracy of downstream applications.

Voiceprint recognition or speaker recognition, is the task of identifying people by their voices. Some of its main subtasks include speaker verification, recognition, diarization, and robust speaker recognition. The purpose of speaker verification is to verify whether a speaker has uttered certain statement based on the prerecorded utterance of the hypothetical speaker. It is usually divided into two stages, as shown in Figure 1. One is the front-end training stage, which is used to extract speaker features and convert utterances in time domain or time-frequency domain into high-dimensional

The associate editor coordinating the review of this manuscript and approving it for publication was Easter Selvan Suviseshamuthu.

feature vectors. Then, in the back-end testing stage, calculate the similarity score between the registered speakers and test speakers, and compare the score with a threshold according to equal error rate (EER).

II. RELATED RESEARCH

Voiceprint recognition techniques have been greatly improved since the rapid development of deep learning and neural networks models. The d-vector model [8], proposed by Google, designed a voiceprint recognition system based on deep neural network (DNN) for the first time, which greatly improves the accuracy compared to the traditional Gaussian models. Later, the x-vector [9] feature based on 3D convolutional neural network (3D-CNN) was developed for voiceprint recognition. Since then, many neural network-based voiceprint recognition model-frameworks have been designed and widely applied. For example, the residual network (ResNet) [10], [11], [12] based on convolutional neural networks, the delayed neural network (TDNN) [13], and the long short-term memory network (LSTM) [14], [15], [16] have all processed voice data from different perspectives to obtain features more conducive to classification and scoring, and model performances are continuously improving.

Voice separation quality can be directly evaluated subjectively or by SNR ratio. It can also be indirectly assessed by the downstream applications such as speech recognition and speaker recognition. The separation quality directly affects the quality of recognition tasks. For example, Settle et al. [17] connected speech separation and speech recognition tasks to construct a multi-task network architecture to prove the practicability of single-channel speech separation. The meeting transcription system proposed by Watanabe et al. [18] provided a solution to the problem of speech overlap in online meetings by using a continuous speech separation method.

In recent years, researchers have also combined the separation task with language separation and speaker recognition tasks. For example, Saeidi et al. [19] combined speech separation with traditional speaker recognition model. Taherian et al. [20] combined single-channel and multi-channel speech separation with DNN/ivector model. Zhao [21] et al., jointly trained the speech separation results based on convolutional recurrent network (CRN) and the speaker verification network. Maciejewski et al. [22] adopted speaker verification as a downstream task of speech recognition and verification of separated results. Aysa et al. [23], [24] combined speech separation with language recognition to improve the accuracy of language recognition in mixed speech through speech separation preprocessing steps.

In this research, we mixed some target and non-target audio according to various power ratios. Voxceleb1 dataset is used for building the training and test sets. In the separation stage, the single channel speech separation model Conv-TasNet is further upgraded to obtain the Conv-TasNet-FS model by adding the feature scaling module. In the speaker verification stage, the MFCCT features are extracted from the MFCC features to improve the time domain characteristics of speech features. Finally, we combine the two stages to reduce EER of the speaker verification task in speech overlapped scenarios.

The rest of this paper is organized as follows. Section III mainly introduces the network architecture and loss function of speech separation, as well as the feature improvement. In Section IV, the dataset, experimental setup, and comparison of experiment results are discussed. The last section is the summary.

III. CONNECTED MODELS

In this section, we introduce the architecture of the combined models of separation and identification. Figure 2 shows the framework of this research. A tandem approach is applied for the single channel overlapped multi-speaker identification task.

A. PRE-TRAINING MODEL OF SPEECH SEPARATION

The single-channel time-domain speech separation models can be designed with three parts, namely encoder, separation network and decoder. Usually, the separation part is the main focus of various mainstream methods. Neural networks models like CNN, RNN, attention mechanism, transformer [25], etc.can be applied to the separation tasks.

The separation model adopted in this paper is Conv-TasNet-FS model, which is named after the addition of α feature scaling module to Conv-TasNet. Conv-TasNet is a time-domain single-channel separation model proposed by Luo and Mesgarani [26]. The encoder part of Conv-TasNet-FS is composed of a 1-dimensional convolution and ReLU activation layer, and the corresponding decoder is its inverse transformation. The separation part is mainly composed of temporal convolution network (TCN). In the 24-layer convolution cycle, the dilated convolution with convolution kernel is used to select features of different ranges.

Inspired by [27], we added a feature scaling block to the 1-D deep expansive convolution block as in Figure 3, which is named α-feature map scaling (α-FMS) block. The FMS is a technique that improves the distinguishability of each feature map in the voice feature extractor. This technology adopted in the research of Hu et al. [28] and Zhang et al. [29] by enhancing the feature map via mixing the information present in each filter. To do this, the FMS method uses a proportion vector, which applies addition, multiplication, or both methods in turn to each filter with values between 0 and 1 to enhance the feature map. Jung et al. [27] improved the FMS technology by incorporating a learnable parameter α, which expands the range from 0 to 1 to the entire real number field, extended the scope of 0 to 1 in FMS. Thus, the added learnable parameters to feature map effectively improved the feature map resolution. As shown in Figure 4 and Figure 5, we add a learned parameter α to the original feature M, then multiply it with S, as shown in formula (1), where S is indirectly obtained from M after going through pooling, fully connected and sigmoid layers, the specific process is shown

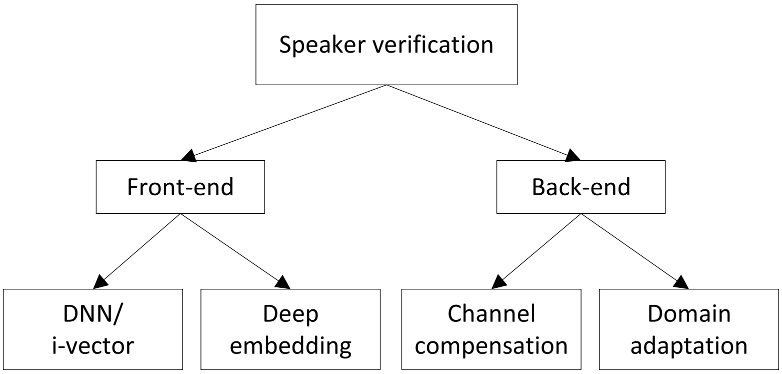

FIGURE 1. The two-stage task of speaker verification.

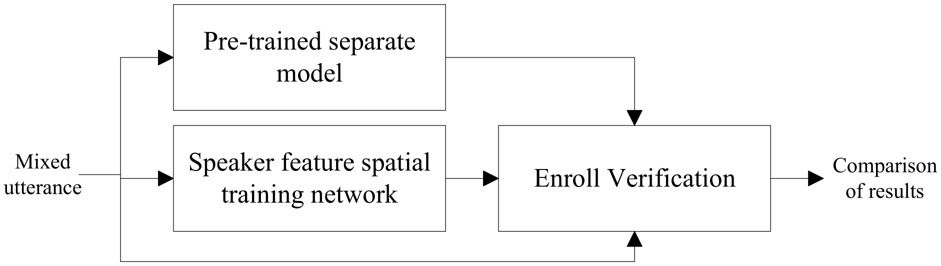
Pre-trained separate model
Mixed utterance
Speaker feature spatial training network
Enroll Verification
Comparison of results

FIGURE 2. Speech separation and speaker recognition framework.

in Figure 4.

$$M' = (M \oplus \alpha) \otimes S \tag{1}$$

B. VOICEPRINT RECOGNITION MODEL

The prototype network model, that is most closely related to our work, is proposed by Chung et al. [11], in which the residual convolutional neural network ResNetSE34 is combined with various loss functions to obtain a metric space through training, and the intra-class and inter-class distances are calculated to perform classification. Therefore, the goal is to increase the distance within different classes and reduce the distance within the same class.

1) FEATURES

In the task of text independent speaker recognition, it is necessary to train the speaker model with a large number of speaker utterances. For each segment of speech, features are extracted through the operations of frame splitting and windowing, and MFCC features are used in this experiment.

For the improvement part, the features fed into the training model are upgraded, that is, on the basis of MFCC, the time-domain features of each frame are extracted to obtain MFCCT features. The specific steps are as follows: (1) Package every 10 rows of each MFCC feature matrix column into a bin; (2) select 10 different time domain parameters as shown in the table from this MFCC bin; (3) write the selected feature parameters into the corresponding columns. The selected parameters are shown in the Table 1:

TABLE 1. Time domain parameters extracted by MFCCT.

Lable	Statistical time-domain features
Min	Minimum value of each bin
Max	Maximum value of each bin
$\mathbf{M}_n$	Mean value of each bin
$\mathbf{M}_d$	Median of each bin
$\mathbf{M}_o$	Mode of each bin
STD	Standard deviation
VAR	Variance of each bin
RMS	Root mean square of each bin
Q	50th percentile of each bin

TABLE 2. Some prototype model architectures. L: Length of the input sequence.

Layer	Kernel size	Stride	Output shape
Conv1	3×3×32	1×1	L×64×32
Res1	3×3×32	1×1	L×34×32
Res2	3×3×64	2×2	L/2×32×64
Res3	3×3×128	2×2	L/4×16×128
Res4	3×3×256	2×2	L/8×8×256
Flatten	-	-	L/8×2048

2) THE TRUNK NETWORK

After that, we apply a thin RESNETse34 model, RESNETse network with input channel 1, for training. As show in Table 2. Then the frame-level information is aggregated into the discourse level embedding through the encoder.

3) CONVERGED NETWORK

As a bridge between the frame layer and the hidden layer of the discourse layer, the aggregation network aggregates speaker features at the frame level into speaker features at

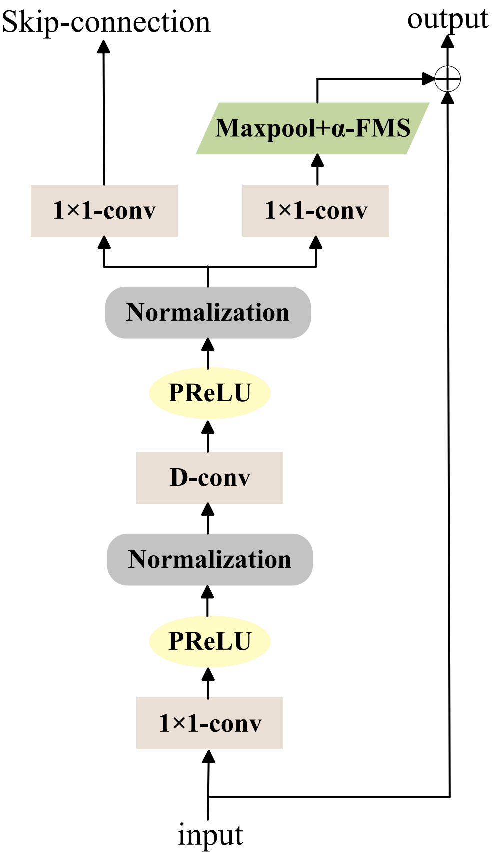

FIGURE 3. 1-D conv-block with α-feature map scaling.

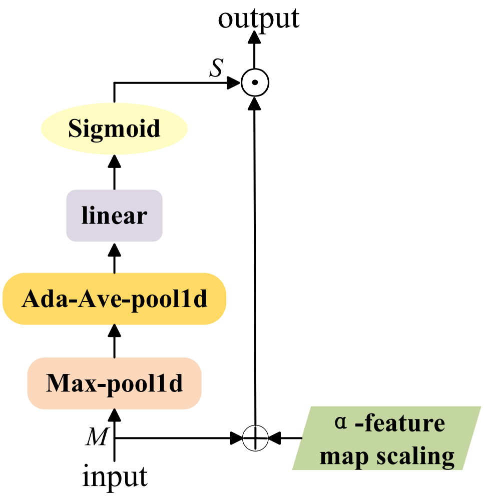

FIGURE 4. The architecture of α-feature map scaling block.

the sentence level. There are two main types of aggregation networks used in this article. One is Self-Attentive pooling (SAP) [30], which focuses on aggregating frame-level features and identifying more formative frames for discourse-level speakers. The other is attentive statistics pooling (ASP) [31] to aggregate time frames and calculate the weighted average and channel weighted standard deviation.

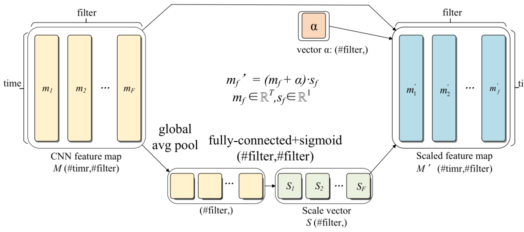

FIGURE 5. Illustration of the α-feature map scaling technique. Independent α is added to each filter of a feature map [27].

- *SAP self-attetion average pooling:*
  the attention mechanism is applied to the pooling layer to compute the importance weight vector so that the neural network can focus on the special parts of the input.
- *ASP attentive statistics pooling:*
  standard deviation is calculated on the basis of self-attention pooling, and the calculation is the same as that of statistics pooling.

4) LOSS FUNCTION

There are many kinds of loss functions in research reports, including Softmax loss and its variants, Pairwise loss, Triplet loss, Quadruplet loss, Prototypical network loss, Angular Prototypical (AP) loss, etc. In this experiment, we choose softmax variants AMsoftmax loss [32] and Angular Prototypical (AP) loss [12] to optimize the metric space.

- *Additive margin softmax (AMsoftmax):*
  different from softmax, which can only divide the boundary between categories, AMsoftmax can reduce the intra-class distance and increase the class spacing, and reduce the class interval to the target region, while generating the class spacing with margin size. The formula is as in (2), as shown at the bottom of the next page. Where τ is a scaling factor for preventing gradients too small during the training process [33]. And m is margin, $\theta_{l_n,n}$ is the angle between the feature vector of the current sample and the feature vector of its real category, and $\theta_{j,n}$ is the angle between the feature vector of the jth category and the feature vector of the current sample.
- *Angular Prototypical:*
  the angular prototype loss is formed using the same batch as the original prototype loss, reserving one utterance from each class as a query. Each of its centroids is made up of the same number of utterances in the support set, so the test scenario can be accurately simulated during training.

IV. THE EXPERIMENTAL SETUP AND DATASET

A. DATASET

Voxceleb1 [34] is a large-scale open source audio and video dataset released by the University of Oxford in 2017, only the audio data is used in this research. The audio and video of Voxceleb1 are taken from YouTube and belong to the real English voice of reality scenarios. Its characteristics include: 1) a wide range of speakers with a variety of accents, occupations and ages; 2) The gender distribution of the dataset is balanced between men and women; 3) The audio sampling rate is 16kHz, 16bit, mono and PCM-WAV format; 4) Speech with a certain amount of real noise, non-artificial white noise, noise appears at irregular time points, the human voice contains loud or weak intervals; 5) Noise includes: environmental burst noise, background human voice, laughter, speech aliasing, echo, indoor noise, recording equipment noise, etc. Voxceleb1 selected a total of 1251 speakers, including 1211 in the training set and 40 in the test set, and these two sets do not overlap. According to the data format required by multi-task, we used the training and testing sets of Voxceleb1 processed in two stages. In the following sections, we describe the data processing method and the data in the separation and identification stages.

1) Data processing method: we apply the same treatment to the training and test sets. In the first step, each audio is aligned and trimmed to the same length of three seconds audio. In the second step, one speaker is selected as the target speaker, such as id10001. While one audio is the target audio, other speakers in the dataset randomly selected as the non-target speakers. We use different power ratios between target speakers and non-target speakers to simulate the far and near distance of two simultaneous speakers in various scenarios. There are 5 ratios are selected, they are [0.5:05, 0.6:0.4, 0.7:0.3, 0.2:0.8, 0.9:0.1]. And only the 100% overlap ratio is used in this research, that means no deviation between mixed voices. Among the power ratio, 0.5:0.5 indicates that the distance between the target speaker and the non-target speaker is equal in the single channel scene, and 0.9:0.1 can indicates that the non-target speaker is relatively distant. Finally, according to the method of the second step, the next target speaker is selected for mixing until all the speakers are recursively selected.
2) Separation stage: the WSJ0-2mix data set is also used for single-channel speech separation, based on Conv-TasNet model in this research. However, [35], [36], [37] and other experiments have proved that the separation model trained on WSJ0-2mix is difficult to adapt to downstream tasks. Therefore, we extracted 14.6G data from the mixed training set of Voxceleb1 at equal intervals, and divided it into 10G training set and 4.6G validation set, used for the original separation model.
3) Recognition stage: In order to verify the effect of speech separation on the voiceprint recognition task, we set up a comparative experiment. Therefore, the datasets of voiceprint recognition are prepared by mix-up with various ratios. The data volume accumulated to 130G for the training set and 5G for the test set. According to the requirements of the verification task, we use the target speaker's clean voice as the registration set to estimate the comparison results between the mixed test set and the separated test set.

B. THE EXPERIMENTAL SETUP

1) SEPARATION EXPERIMENT

The speech separation pre-training model adopts the same experimental configuration as [26] and adopts the network form of encoder-separator-decoder. The difference is that we only retain the target speaker's speech after separation and for the downstream voiceprint recognition task. The experiment is based on pytorch 1.12.0 framework and trained on NVIDIA GeForce RTX 3090 GPU with 24G memory. The model optimizer uses Adam with an initial learning rate of 0.001. The batch size is set to 24, and the total training time on a single card is about 33 hours.

2) RECOGNITION EXPERIMENT

The experiment was conducted on a Quadro RTX A5000 GPU with 16G video memory based on pytorch 1.12.0 framework. The maximum epoch is set to 100, and the batch size is set to 200, that is, each batch reads one speech of 200 speakers. The optimizer uses Adam with an initial learning rate set to 0.001 and drops by 25% after every 3 epochs. Referring to [11] we only selected some parameters for comparative experiment in the loss function, including: AM-softmax selected margin=0.2, scale=30 and margin=0.3, scale=30. The equal error rate (EER) of the evaluation index refers to the value when the false acceptance rate (FAR) and the false rejection rate (FRR) are equal. A smaller value of the indicator indicates better performance. We trained 100 epochs in about 3.3 days per experiment and spend about 4 days after adding MFCCT features.

C. RESULTS

1) SPEECH SEPARATION EXPERIMENT

For separated results, we use the classic speech separation evaluation indicators signal-to-distortion ratio improvement (SDRi) and scale invariant signal-to-noise ratio (SI-SNR).

The SDRi index is defined as the difference between the SDR (Signal-to-Distortion Ratio) of the separated speech signal and the SDR of the mixed speech signal, both measured

$$\mathcal{L}_{\mathrm{AMS}} = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\exp\left(\tau\left(\cos\left(\theta_{l_n,n}\right)-m\right)\right)}{\exp\left(\tau\left(\cos\left(\theta_{l_n,n}\right)-m\right)\right)+\sum_{j=1,j\neq l_n}^{J}\exp\left(\tau\left(\cos\left(\theta_{j,n}\right)\right)\right)} \tag{2}$$

TABLE 3. Comparison witn other methods on WSJ0-2mix dataset.

	SDRi	SI-SNR	Model size
DPCL++ [2]	-	10.8	13.6M
uPIT-BLSTM-ST [3]	10.0	-	92.7M
Conv-TasNet [26]	14.59	14.29	5.1M
Conv-TasNet-FS	**15.07**	**14.59**	5.1M

in decibels (dB):

$$SDRi = SDR(separated) - SDR(mixed) \tag{3}$$

SI-SNR measures the similarity between the estimated source signal and the true source signal in terms of their power ratio, and it is defined as the ratio of the power of the true source signal to the power of the residual noise after separation. Its calculation formula is as follows:

$$\boldsymbol{s}_{\text{target}} = \frac{<\hat{\boldsymbol{s}}, \boldsymbol{s}> \boldsymbol{s}}{\|\boldsymbol{s}\|^2} \tag{4}$$

$$\boldsymbol{e}_{\text{noise}} = \hat{\boldsymbol{s}} - \boldsymbol{s}_{\text{target}} \tag{5}$$

$$SI - SNR = 10\log_{10}\frac{\|\boldsymbol{s}_{\text{target}}\|^2}{\|\boldsymbol{e}_{\text{noisis}}\|^2} \tag{6}$$

Where $\hat{s} \in \mathbb{R}^{1\times T}$ and $s \in \mathbb{R}^{1\times T}$ are the estimated and original clean sources, respectively.

In order to maintain the generality, WSJ0-2mix was used to verify the training and test of the improved network of 1-dimensional depth expansion convolutional blocks, and the results were compared with those before the improvement. Meanwhile, we compared the classic neural network separation models, DPCL++ [2] and uPIT [3] models, based on the WSJ0-2mix dataset. DPCL++ is an end-to-end presentation of the deep cluster method, which mainly converts speech signals into a high-dimensional feature space and then clusters the signals in that space. Utterance-level permutation invariant training (uPIT), as its name suggests, uses an utterance-level cost function to eliminate the need to solve an additional permutation problem during inference. The results are shown in Table3:

2) COMPARISON BEFORE AND AFTER SEPARATION

In order to verify the effectiveness of voice separation on voiceprint recognition, the mixed Voxceleb1 training set was used as the training set of voiceprint recognition model to train the embedding space, and the mixed test set is compared with the separated test set of Conv-TasNet-FS separation network. The results are shown in Table 4.

We estimated the mixture of five different power ratios on the effectiveness of speaker verification results after the preprocessing step, as shown in Figures 6-9. From the results we can see that when the power ratio of the target speaker to the non-target speaker is 0.5:0.5, proposed method demonstrated significant improvement for the separated speech. And the higher target power ratio in the overlapped speech is beneficial for the overall performance. Among them, SAP-AM, SAP-AP, ASP-AM, and ASP-AP respectively selected SAP and SAP for conversion, and trained the models through training. AM only selected the option with better performance, m = 0.2 and s = 30 in parameter selection.

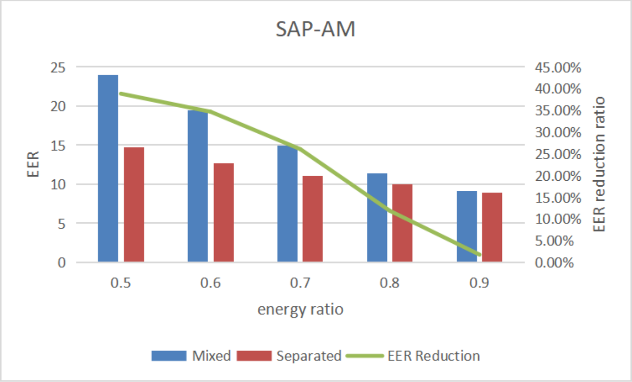

FIGURE 6. Validation results of SAP-AM trained speaker recognition model on mixed and separated data of five power ratios.

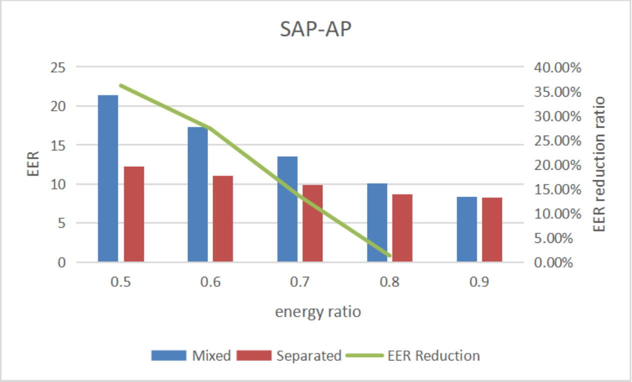

FIGURE 7. Validation results of SAP-AP trained speaker recognition model on mixed and separated data of five power ratios.

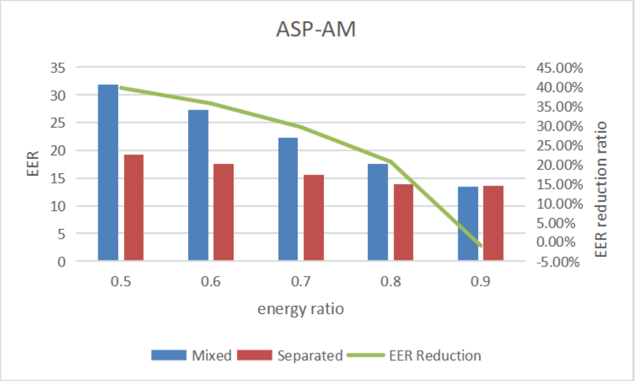

FIGURE 8. Validation results of ASP-AM trained speaker recognition model on mixed and separated data of five power ratios.

3) IMPROVED FEATURES BEFORE AND AFTER COMPARISON

In terms of feature extraction, we use MFCCT to extract time domain parameters from MFCC. In [38], MFCCT improves the accuracy by about 50% in the gender recognition

TABLE 4. Multitask - Separate – recognition test results.

Converged	Loss function		Mixed EER(%)	Separated EER(%)	Reduction rate(%)
SAP	AM	m=0.2,s=30	16.063	11.482	28.52
		m=0.3,s=30	16.377	11.797	27.97
	AP		14.863	10.131	**31.84**
ASP	AM	m=0.2,s=30	22.325	15.438	30.85
		m=0.3,s=30	22.630	15.744	30.43
	AP		21.239	16.364	22.95

Some abbreviations: SAP = Self-attentive pooling, ASP = Attentive Statistics Pooling, AM = Additive margin softmax, AP = Angular Prototypical, m = margin, s = scale. **Mixed** represents mixed data of all proportions, and **Separated** represents data after separating all mixed data

TABLE 5. Results of comparison between MFCC and MFCCT.

Feature	Loss function	Mixed EER(%)	Separated EER(%)	Reduction rate(%)
MFCC	AM	16.063	11.482	28.52
	AP	14.863	10.131	31.84
MFCCT	AM	15.896	11.235	29.32
	AP	**14.432**	**9.638**	**33.22**

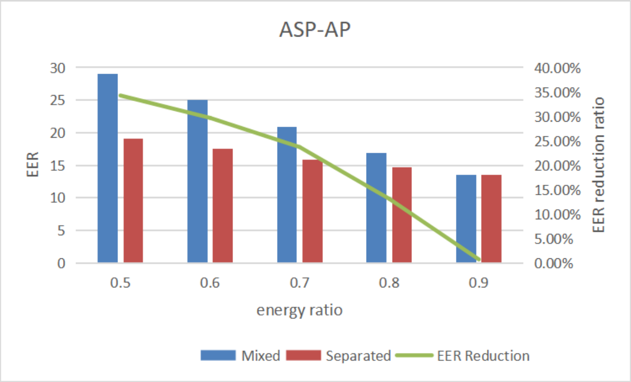

FIGURE 9. Validation results of ASP-AP trained speaker recognition model on mixed and separated data of five power ratios.

experiment. Due to limited time and equipment, the experiment with improved features only selects the aggregation network SAP with the loss functions AMsoftmax (margin=0.2, scale=30) and Angular Prototypical. The comparison results are in Table 5.

V. SUMMARY

In order to solve the problem of target speaker verification in a multi-speaker scenario, we propose a combination of the two tasks. The target speaker speech is obtained through the separation phase, and the target speaker identity is verified in the recognition phase. In the separation phase, we propose the Conv-TasNet-FS method of adding feature scaling module and verified its effectiveness. In the recognition stage, we added MFCCT features and verified its effectiveness through experiments. The experimental results show that the EER of the separated speaker verification tasks has been significantly reduced. In the future work, we will continue to research the speaker verification task in the multi-speaker scenarios and try to integrate the training steps into the overall model framework.

REFERENCES

[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., speech signal Process. (ICASSP)*, May 2016, pp. 31–35.

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

[3] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[4] Y. Luo and 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)*, Apr. 2018, pp. 696–700.

[5] S. Lutati, E. Nachmani, and L. Wolf, "SepIt: Approaching a single channel speech separation bound," 2022, *arXiv:2205.11801*.

[6] X. Yang and C. Bao, "Embedding recurrent layers with dual-path strategy in a variant of convolutional network for speaker-independent speech separation," 2022, *arXiv:2203.13574*.

[7] R. Paturi, S. Srinivasan, K. Kirchhoff, and D. Garcia-Romero, "Directed speech separation for automatic speech recognition of long form conversational speech," 2021, *arXiv:2112.05863*.

[8] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2017, pp. 999–1003.

[9] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333.

[10] L. Jiahong, B. Jie, C. Yingshuang, and L. Chun, "An adaptive ResNet based speaker recognition in radio communication," in *Proc. IEEE Int. Conf. Emergency Sci. Inf. Technol. (ICESIT)*, Nov. 2021, pp. 161–164.

[11] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," 2020, *arXiv:2003.11982*.

[12] H. Soo Heo, B.-J. Lee, J. Huh, and J. Son Chung, "Clova baseline system for the VoxCeleb speaker recognition challenge 2020," 2020, *arXiv:2009.14153*.

[13] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," 2020, *arXiv:2005.07143*.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in *Supervised Sequence Labelling With Recurrent Neural Networks*. Springer, 2012, pp. 37–45.

[15] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with LSTM," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5239–5243.

[16] Z. Zhao, H. Duan, G. Min, Y. Wu, Z. Huang, X. Zhuang, H. Xi, and M. Fu, "A lighten CNN-LSTM model for speaker verification on embedded devices," *Future Gener. Comput. Syst.*, vol. 100, pp. 751–758, Nov. 2019.

[17] S. Settle, J. L. Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-end multi-speaker speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4819–4823.

[18] S. Watanabe, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," 2020, *arXiv:2004.09249*.

[19] R. Saeidi, P. Mowlaee, T. Kinnunen, Z.-H. Tan, M. G. Christensen, S. H. Jensen, and P. Fränti, "Signal-to-signal ratio independent speaker identification for co-channel speech signals," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 4565–4568.

[20] H. Taherian, Z. Wang, J. Chang, and D. Wang, "Robust speaker recognition based on single-channel and multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 28, pp. 1293–1302, 2020.

[21] F. Zhao, H. Li, and X. Zhang, "A robust text-independent speaker verification method based on speech separation and deep speaker," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6101–6105.

[22] M. Maciejewski, S. Watanabe, and S. Khudanpur, "Speaker verification-based evaluation of single-channel speech separation," in *Proc. Interspeech*, Aug. 2021, pp. 3520–3524.

[23] Z. Aysa, M. Ablimit, H. Yilahun, and A. Hamdulla, "Language identification-based evaluation of single channel speech separation of overlapped speeches," *Information*, vol. 13, no. 10, p. 492, Oct. 2022.

[24] Z. Aysa, M. Ablimit, and A. Hamdulla, "Multi-scale feature learning for language identification of overlapped speech," *Appl. Sci.*, vol. 13, no. 7, p. 4235, Mar. 2023.

[25] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 21–25.

[26] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[27] J.-W. Jung, H.-J. Shim, J.-H. Kim, and H.-J. Yu, "α-feature map scaling for raw waveform speaker verification," *J. Acoust. Soc. Korea*, vol. 39, no. 5, pp. 441–446, 2020.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[29] J. Zhang, N. Inoue, and K. Shinoda, "I-vector transformation using conditional generative adversarial networks for short utterance speaker verification," 2018, *arXiv:1804.00290*.

[30] G. Bhattacharya, M. J. Alam, V. Gupta, and P. Kenny, "Deeply fused speaker embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2018, pp. 3588–3592.

[31] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," 2018, *arXiv:1803.10963*.

[32] M. Hajibabaei and D. Dai, "Unified hypersphere embedding for speaker recognition," 2018, *arXiv:1807.08312*.

[33] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2019, pp. 1652–1656.

[34] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," 2017, *arXiv:1706.08612*.

[35] B. Kadioglu, M. Horgan, X. Liu, J. Pons, D. Darcy, and V. Kumar, "An empirical study of conv-TasNet," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7264–7268.

[36] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A dissecting approach," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6359–6363.

[37] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," 2020, *arXiv:2005.11262*.

[38] R. Jahangir, Y. W. Teh, N. A. Memon, G. Mujtaba, M. Zareei, U. Ishtiaq, M. Z. Akhtar, and I. Ali, "Text-independent speaker identification through feature fusion and deep neural network," *IEEE Access*, vol. 8, pp. 32187–32202, 2020.

RONG JIN is currently pursuing the master's degree with the School of Information Science and Engineering, Xinjiang University. Her current research interests include speech separation and speaker verification.

MIJIT ABLIMIT received the Ph.D. degree from Kyoto University, Japan, in 2012. He is currently a Professor with the School of Information Science and Engineering, Xinjiang University. His current research interests include speech and language information processing, audio retrieval, speech dialogue, information retrieval, and robot and intelligent information processing.

ASKAR HAMDULLA received the Ph.D. degree from the University of Electronic Science and Technology of China, in 2003. He is currently a Professor with the School of Information Science and Engineering, Xinjiang University. His current research interests include speech recognition and synthesis, pattern recognition and image processing, natural language processing, information retrieval, and content security.