Electronic Engineering and Informatics
G. Izat Rashed (Ed.)

doi:10.3233/ATDE240068

The Speaker Verification System Based on TDNN Improvements in Noisy Environments

Qingyi HE[a1], Qingning ZENG[a2], Xuejun ZHAO[b3]
[a]*School of Information and Communication, Guilin University of Electronic Technology, Guilin, China*
[b]*School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin, China*

Abstract. The rise of deep learning technology has significantly improved the recognition rate of voiceprint recognition technology, such as the success of the X-vector architecture, which utilizes Time Delay Neural Networks (TDNN) to transform variable-length speech segments into fixed-length outputs. However, the current popular voiceprint recognition models have significantly decreased applicability in noisy environments. To address this issue, this study investigates the limitations of the X-vector architecture and proposes an improved speaker verification model based on TDNN. This model incorporates Long Short-Term Memory (LSTM) to model the input speech features while retaining information related to previous time steps. Similar to the ECAPA-TDNN model, we introduce a one-dimensional Res2Net module with a channel attention mechanism (SE-Res2Block) at the frame level, which enhances channel correlation and rescales channels based on recorded global properties, thereby extending the temporal context of the frame layer. Finally, the model's feature representation capacity is enhanced through multi-layer aggregation. The results show that the recognition performance of this system reaches 96.32% in a 15 dB noise environment. Furthermore, this system outperforms the commonly used ECAPA TDNN model, demonstrating good accuracy and robustness.

Keywords. Speaker Verification; TDNN; LSTM; Channel Attention Mechanism; Multi-layer Aggregation; X-vector;

1. Introduction

Voiceprint is a spectrum of sound waves that carries speech information. Voiceprints not only possess uniqueness but also exhibit relative stability, as an individual's voice tends to remain unchanged for a long time after adulthood. Therefore, similar to fingerprints and irises, voiceprints can be used as a unique biological feature for identity verification. Compared to traditional fingerprint and iris recognition, voiceprint recognition

[1] Qingyi HE, School of Information and Communication, Guilin University of Electronic Technology, e-mail: 1379521566@qq.com

[2] Qingning ZENG, School of Information and Communication, Guilin University of Electronic Technology, e-mail: qingningzeng@126.com

[3] Corresponding author: Xuejun ZHAO, School of Electronic Engineering and Automation, Guilin University of Electronic Technology, e-mail: 470610369@qq.com

technology offers advantages such as non-contact operation, easy collection, high efficiency, and high accuracy. As a result, voiceprint recognition finds extensive applications in various fields, including security authentication, telephone banking, voice assistants, and more.

Voiceprint recognition can be divided into speaker identification and speaker verification [1]. Speaker identification aims to determine which person among several individuals uttered a particular speech segment involving a "multiple-choice" problem. On the other hand, speaker verification confirms whether a given speech segment belongs to a specific designated individual, involving a "one-to-one discrimination" problem. Despite the widespread application of voiceprint recognition technology, real-world environments pose challenges such as noise, echoes, and interference, which can degrade the original audio quality and even obscure the characteristic speaker information. Therefore, it is crucial to reduce the factors that cause interference in voiceprint signals to enhance the robust performance of speaker verification systems.

The concept of "voice print" was born in Bell Laboratory in 1945. L.g. Kesta et al. proposed a Spectrogram matching problem [1]. Voiceprint recognition is mainly divided into two parts: feature extraction and speaker modeling. Traditional voice print recognition methods are mainly template matching and statistical analysis algorithms [2]. At present, the mainstream voice print recognition technologies mainly include the GMM-UBM model, i-vector model, d-vector model, and X-vector model [3-7]. In addition, the use of time-delay neural networks to model speaker features is currently a very active research area [8,9,10]. For example, some researchers enhanced the frame-level feature extraction capability by introducing one-dimensional SE-Res2Block based on TDNN [11]. The author added LSTM based on TDNN to better capture the time information in speech [12,13]. This paper proposes a speaker confirmation model based on TDNN and LSTM, which incorporates a channel attention mechanism and multi-layer fusion mechanism and will be explained in detail in Section 2. In addition, the SpecAugment algorithm was added as front-end noise reduction to improve the robustness of the system and enhance the feature expression ability.

2. Network Design

2.1 Network improvement

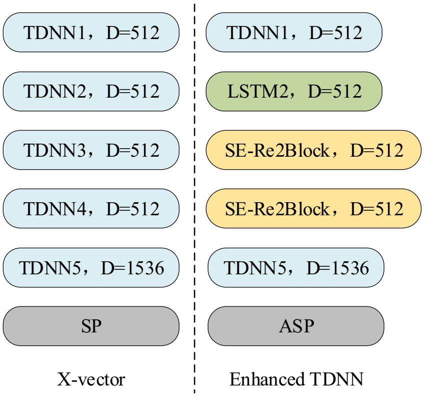

Figure 1. The network architecture of TDNN for X-vectors and Enhanced TDNN

We have made improvements to the original X-vector based on TDNN, as shown in Figure 1. The first part replaces a TDNN layer with an LSTM. The second part replaces two traditional TDNN layers with two one-dimensional SE-Res2Blocks. The third part utilizes Attentive Statistic Pooling (ASP) [14] with attention instead of a single Statistic Pooling (SP) layer.

2.2 Integral design

Figure 2 represents the schematic diagram of the overall design of the network model in this paper. It mainly consists of six parts: TDNN+ReLU+BN module, LSTM module, two layers of one-dimensional SE-Res2Block modules, TDNN+ReLU module, ASP+BN module, and AAM-Softmax module.

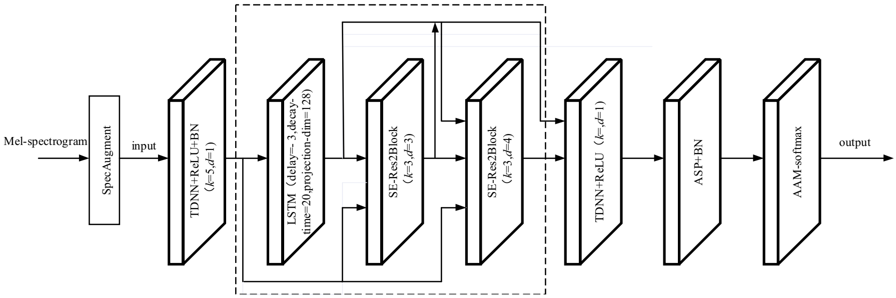

Figure 2. The model in this paper

In the figure, ReLU is a nonlinear activation function. BN is batch normalization [15]. FC is the fully connected layer. *k* is the kernel size, and *d* is the expansion spacing of the TDNN layer or SE-Res2Blocks.

Firstly, the Mel-spectrogram processed by SpecAugment is inputted into the TDNN layer. This layer performs convolution operations along the time axis using multiple convolutional kernels to extract feature vectors at each time point. These feature vectors are then combined to form a new sequence of feature vectors.

A speech signal is a temporal signal, where each time point is correlated with the preceding and subsequent time points. Therefore, to better capture the temporal information in the speech signal, a neural network structure capable of processing sequential data needs to be used. To achieve this, this article introduces the LSTM layer.

The LSTM layer is a type of recurrent neural network structure that has memory cells and gate structures and can effectively process sequential data. In this article, the LSTM layer receives the feature vector sequence extracted by the TDNN layer and models it. By reusing the same neurons and introducing three gate structures (input gate, forget gate, and output gate), the LSTM layer can effectively solve problems such as gradient vanishing and gradient explosion and model long-term dependencies in sequential data. In addition, LSTM has a certain degree of noise resistance and can effectively distinguish target information and noise information in speech signals by learning the temporal patterns and context of input data, thereby improving the accuracy and robustness of the model.

In the original X-vector system, the time context of frame layers is limited to 15 frames. It is indicated that SE-Res2Blocks are beneficial for enhancing the depth and nonlinearity of the model [10]. To enhance the representation and generalization abilities of the model, this paper introduces two layers of one-dimensional SE-Res2Blocks after the LSTM layer, which enables the model to better capture the temporal information and frequency domain features of speech signals. SE-Res2Blocks combines the characteristics of SE modules and Res2Blocks, where the SE module is used to adaptively adjust the importance of different feature channels to enhance the expressive power of feature channels. Res2Blocks is a novel residual block composed of multiple ResBlocks and a 1D convolution layer, which is used to improve the model's non-linear modeling ability. In addition, this paper adopts a multi-layer aggregation mechanism, where the output features of each module can serve as the input features of the next module, to extract more advanced features layer by layer. This multi-layer feature aggregation can help the model better capture the temporal and frequency characteristics of speaker signals and further improve the performance of speaker recognition.

Subsequently, the outputs of LSTM and SE-Res2Block are input together into the TDNN+ReLU module for aggregation and then fed into the ASP layer to obtain a global speaker embedding. Then, the BN layer maps the global speaker embedding to a low-dimensional vector space, further enhancing the expressive power of the features. Finally, the feature vector is classified using AAM-Softmax to determine the speaker identity corresponding to the input speaker signal.

2.3 Data augmentation

Before inputting the speaker embedding into the model, the SpecAugment algorithm is applied to the Mel-spectrogram, which includes two parts: time masking and frequency masking. Specifically, some contiguous frequency regions are selected on the time axis and frequency axis of the Mel-spectrogram, and the spectral values of these regions are set to zero. This method can simulate the speech variations and discontinuities in the real world, as well as the frequency variations and discontinuities in the signal, to provide more robust training data that helps improve the generalization ability of the model and reduce overfitting.

3. Experiment and Analysis

3.1 Data set

This paper uses the Chinese speech dataset (zhvioce) for model training. The dataset has been processed for noise reduction and silence removal and contains 3, 200 speakers, with a total audio duration of about 900 hours and approximately 130 million words. 80% of the dataset is used for training, and the rest is used for testing. In addition to the aforementioned dataset, the VoxCeleb1-E and VoxCeleb1-H datasets are also used for testing. These two datasets are widely recognized as large-scale standard datasets for speaker recognition. The former contains 6, 366 speakers, and the latter contains 13, 718 speakers, allowing the performance of the model to be evaluated in different testing environments.

3.2 Experimental procedure

The experiments in this paper are divided into two stages: training and testing, as shown in Figure 3. In the training stage, a TDNN-improved neural network is used to construct the speaker identity model database. In the testing stage, after preprocessing the test speech, the speaker embedding parameters are extracted and inputted into the model. The scores are calculated by comparing them with the model database, and the results are output.

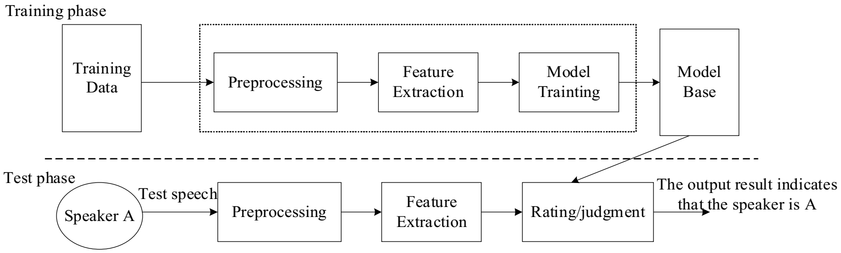

Figure 3. Experimental flowchart

In the experiments, in the training stage, the audio in the training dataset is first preprocessed by using Voice Activity Detection (VAD) technology to clip the speech signal, remove the silent parts, and filter out the dataset with a speech length of less than 0.5 s. Then, the speech dataset is resampled and normalized to a 20 dB gain. The length of the training speech is set to 3 s, and the audio, shorter than the training length, is padded through copying. The speech signal is pre-emphasized by using a first-order high-pass filter, and then the fast Fourier transform is used to obtain the spectrogram after framing. A Hamming window is applied to each frame signal after framing to obtain a better sidelobe attenuation. Finally, the Mel-spectrogram is obtained by using the dot product of the power spectrum and the Mel filter bank. During training, each audio segment is represented by the corresponding Mel-spectrogram, and each frequency band corresponding to each frame is a 513-dimensional feature. The optimizer used for the model is Adam, and the loss function is the Cross-Entropy Loss Function. The initial learning rate is set to 0.001, and a total of 30 training epochs are set, with about 3242 audio segments per epoch. Table 1 shows the experimental results of the model in the testing datasets of zhvioce, VoxCeleb1-E, and VoxCeleb1-H.

Table 1. The recognition rate of the system in different test sets in this paper

Data set	Tpr	Fpr	Eer
zhvoice	0.98972	0.00730	0.01758
VoxCeleb1-E	0.97881	0.00788	0.02907
VoxCeleb1-H	0.98342	0.00776	0.02434

3.3 Comparison experiment

To verify the feasibility of the speaker verification system proposed in this paper under different signal-to-noise ratio (SNR) noise environments, the system is compared with the TDNN model and the ECAPA TDNN system in speaker verification experiments, and the recognition rates are calculated for these three systems. The results are shown in Table 2.

Table 2. Experimental results in different noise environments

Types of noise	Environment	TDNN	ECAPA TDNN	Textual algorithm
		EER%	EER%	EER%
White	Noise-free	3.7802	3.1894	2.7392
	5 dB	17.9766	15.7944	14.9590
	10 dB	7.1595	6.8948	6.2648
	15 dB	5.4102	4.9363	3.5199
Factory	Noise-free	3.1802	3.5987	2.6135
	5 dB	17.4075	16.5883	15.0262
	10 dB	7.3496	6.8952	6.3266
	15 dB	5.5851	5.0423	3.6779

The following conclusions can be drawn from the experimental results of the three algorithms under white noise and factory noise environments. In a noise-free environment, all three algorithms show a high recognition rate. Among them, the text algorithm achieves the highest recognition rate of 97.26% compared to the other two algorithms. In a noisy environment, as the signal-to-noise ratio (SNR) increases, the recognition rates of all models improve, indicating that the noise environment has a significant impact on the recognition rate of the speaker verification system. In the case of low SNR, the proposed algorithm performs slightly better than the other two algorithms, indicating that the proposed system has a relatively strong noise resistance performance. In a 15 dB white noise environment, the proposed algorithm improves the system's recognition rate to over 96%, with a relative error rate reduction of 34.93% compared to the TDNN network and 28.69% compared to the ECAPA TDNN network. Although the recognition rate of the proposed algorithm slightly decreases in 5 dB and 10 dB noise conditions, the overall performance of the proposed system in noise environments is significantly improved, demonstrating the robustness of the proposed speaker verification system in noisy environments.

4. Conclusion

To improve the accuracy of the speaker verification system in noisy environments, this paper proposes a speaker verification model based on an improved TDNN. The model has the following four key features: firstly, the specAugment algorithm is introduced for data augmentation to enhance the model's noise resistance and robustness; LSTM is introduced to effectively capture long-term dependencies in speech signals and provide more comprehensive contextual information for better modeling of speaker characteristics; a channel attention module is added after each convolutional layer to allow the network to focus more on voiceprint features that have not been discovered or activated at the same or similar time nodes, thus strengthening the importance of voice features corresponding to each channel; a novel cross-layer propagation and aggregation mechanism is used to promote information exchange between different layers and improve model performance. Experimental results show that the proposed system achieves a recognition rate of 97.26% in a noise-free environment and over 96% in a 15 dB noisy environment. In addition, in a 15 dB noisy environment, the proposed model outperforms the TDNN model by 1.9% and the ECAPA TDNN model by 1.4%.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 61961009), and the Foundation of Guangxi Wireless Broadband Communication and Signal Processing Key Laboratory (Grant No. GXKL06200107). Innovation Project of Guangxi Graduate Education (Grant No. 2022YCXS042).

References

[1] Zeng C, Ma C F, Wang Z F, Zhu D L, Zhao N, Wang J Liu C 2022 *A Review of Speaker Recognition Research under the Framework of Deep Learning* (China: Computer engineering and Application vol 7) pp 8-16

[2] Lu Y N, Shan B V, Guan C 2017 *The current status and development application of voiceprint recognition technology* (China: Information Systems Engineering vol 2), p 11.

[3] Zheng C J, Wang C L, Jia N 2020 *Overview of Acoustic Feature Extraction in Speech Tasks* (China: Computer Science vol 5) pp 110-119.

[4] Wang Y Q, Zhang W, Yang B Y 2020 *Speaker recognition based on adaptive Gaussian Mixture model in noisy environment* (China: Technology Perspective vol 17), pp 46-47.

[5] M. Li, A. Tsiartas, M. Van Segbroeck, and S. S. Narayanan, 2013 *Speaker verification using simplified and supervised i-vector modeling* (Canada: IEEE International Conference on Acoustics, Speech and Signal Processing) pp 7199-7203.

[6] KANAGASUNDARAM A, SRIDHARAN S, GANAPATHYS 2019 *A study of x -vector based speaker recognition on short utterances* (Proceedings of the Interspeech) pp 2943-2947.

[7] LIU J, PI J, XIA L 2020 *A novel and high precision tomato maturity recognition algorithm based on multi-level deep residual network* (China: Multimedia Tools and Applications vol 3) pp 1-15.

[8] Chen C B Zhao L 2010 *Speaker Verification using GMM-UBM with Embedded TDNN.* (Signal Processing vol 4) pp 564-568.

[9] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur 2019 *Speaker recognition for multi-speaker conversations using x-vectors* (Proc ICASSP*)*, pp 5796–5800.

[10] D. Garcia-Romero, A. McCree, D. Snyder, and G. Sell 2020 *JHUHLTCOE system for the VoxSRC speaker recognition challenge* (Proc. ICASSP), pp 7559–7563.

[11] DESPLANQUES B, THIENPONDT J, DEMUYNCK K. 2020 *Ecapa-tdnn: emphasized channel attention, propagation and aggregation in tdnn based speaker verification.* (Belgium: Proceedings of the Interspeech) pp 3830-3834.

[12] C.-P. Chen, S.-Y. Zhang, C.-T. Yeh, J.-C. Wang, T. Wang, and C.-L. Huang 2019 *Speaker characterization using tdnn-lstm based speaker embedding in ICASSP* (IEEE International Conference on Acoustics, Speech and Signal Processing), pp.6211–6215.

[13] Jin H, Zhu W B, Duan Z K 2021 *TDNN-LSTM model based on attention mechanism and its application* (Acoustic Technology vol 04) pp 508-514.

[14] Okabe K, Koshinaka T, Shinoda K 2018 *Attentive Statistics Pooling for Deep Speaker Embedding.* (Interspeech)p 993.

[15] S. Ioffe and C. Szegedy K 2015 Batch normalization: Accelerating deep network training by reducing internal covariate shift (Proc.ICML)pp. 448–456.