Journal of Electronic Research and Application, 2026, Volume 10, Issue 3
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

A Tibetan Speaker Verification Method Based on the Improved MFA-NConformer Model

Yitong Gong*, Yuting Chen

Xinjiang Vocational University of Technology, Xinjiang, China

***Corresponding author:** *Yitong Gong, 2030642289@qq.com*

Abstract: MFA-conformer methods are widely used in English and Chinese speaker recognition. Theoretically language-independent but practically language-related, Tibetan speaker recognition currently relies on traditional models with poor performance. To address this, we adopt MFA-conformer as the basic framework and propose improvements: integrating 1D depth-wise separable convolution and channel attention into the conformer feed-forward network, fusing multi-block features, and adding an intra-class correlation regularizer to GE2E loss. Experiments show the improved model reduces the equal error rate (EER) compared with the conformer baseline.

Keywords: Conformer block; Tibetan; GE2E loss; Speaker verification

Online publication: April 24, 2026

1. Introduction

Speaker recognition technology achieves automated identity verification by analyzing voice parameters, representing a key breakthrough in the security field [1]. Since the rise of deep learning, significant progress has been made in speaker verification technology. A keyword-based text relevance speaker verification system has explored various attention mechanisms. Research indicates that self-attention models can effectively integrate relevant information within input sequences [2]. As a revolutionary new approach, the transformer model proposes an innovative solution that disrupts traditional recurrent and convolutional layer structures [3]. The transformer relies entirely on attention mechanisms, replacing the recurrent layers in traditional encoder-decoder architectures with multi-head self-attention. This self-attention mechanism serves as the core component of the transformer, enabling powerful modeling of global interactions between speech frames [4,5]. However, global self-attention still appears to be insufficient in capturing local information that is critical for speaker embedding. To address this problem, convolutional neural networks (CNNs) were introduced. A study have proposed the conformer model, which cleverly fuses components of CNN and transformer to achieve end-to-end speech recognition with the ability to extract both local and global contextual information [6]. In addition, another study further proposes

the multi-scale feature aggregation conformer (MFA-Conformer), which is a simple and effective backbone for automated speaker verification based on a convolutionally enhanced transformer (conformer) [7]. MFA-conformer is able to capture global and local features more efficiently by connecting the frame-level outputs of all conformer blocks and achieves more significant improvements in speaker feature extraction compared to traditional conformers. Speaker verification technology has progressed in the era of deep learning by introducing end-to-end modeling, transformer, conformer, and other techniques. Several studies have conducted cross-language speaker verification studies on the English-Mandarin-Viennese dialect dataset and the English-Mandarin-Taiwanese dialect dataset, respectively [8,9]. Theoretically, speaker recognition technology captures acoustic features independent of language or content. However, experimental results indicate that matching rates remain low when the same speaker registers and tests using different languages [10]. Proposed modeling spoken language or dialects as speaker features through neural networks, and completed verification based on this approach. This study employs a Tibetan corpus for speaker recognition tasks. Tibetan belongs to the Tibeto-Burman language family and is classified as a low-resource language due to its unique phonological and syntactic structures [11,12]. This paper aims to extract more individualized Tibetan speaker features using Tibetan corpus, get better validation results, and propose an improved model by exploring the structure of the MFA-conformer model, multi-scale feature aggregation new conformer (MFA-NConformer).

2. Research methodology

2.1. Tibetan speaker verification system framework

Figure 1 shows the framework for Tibetan speaker verification, which consists of training and testing phases. The core goal is to verify whether an audio sample comes from the target speaker. In the training phase, acoustic features are extracted from enrolled speakers' audio data and input into a neural network for training to establish a voiceprint database. In the testing phase, acoustic features are extracted from the test speaker's audio data, and the extracted voiceprint features are matched against those in the database to confirm the speaker's identity.

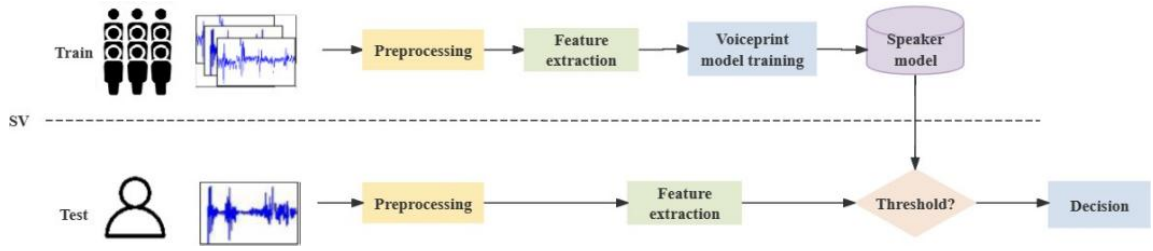

Figure 1. Tibetan speaker recognition system framework.

2.2. Loss function

2.1.1. GE2E loss

Generalized end-to-end loss (GE2E loss) is a loss function for speaker recognition tasks [13]. It maps speech feature vectors to a speaker-specific embedding space during training and leverages the Euclidean distance between embedding vectors for speaker verification. The optimization goal is to minimize intra-class distances and maximize inter-class distances. Its calculation involves the Euclidean distance comparison and a contrastive loss function to pull positive sample pairs closer and push negative sample pairs apart, thereby optimizing speaker recognition performance. The mathematical expression is given as follows:

$$L_{GE2E} = -\frac{1}{N}\sum_{j,i} log \frac{e^{S_{j,i,j}}}{\sum_{k=1}^{N} e^{S_{j,i,k}}} \quad (1)$$

Where GE2E loss creates a phase matrix that defines the cosine similarity between each embedding and all centers of mass:

$$S_{j,i,k} = \omega \cdot cos\,(x_{j,i}, c_k) + b \quad (2)$$

ω and b are learnable scales and biases.

2.1.2. GE2Eicr loss

The "Intra-class Correlation Regularizer" (ICR) introduces the concept of intra-class sample similarity into the loss function. Through increasing the similarity among samples within the same class, it enhances the model's ability to distinguish between samples of the same category.

The formula for the ICR regularization term is as follows:

$$L_{ICR} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}^{N} cos\,(s_{ii}, s_{ij}) \quad (3)$$

Where cos(*) denotes the cosine similarity, s_{ii} denotes the feature vector (embedding vector) of the i sample within the same category; and s_{ij} denotes the feature vector of the i sample and the j sample within the same category.

In combining the effects of both GE2E Loss and ICR, we define the overall loss:

$$L_{\text{total}} = L_{GE2E} + \omega_{ICR} \times L_{ICR} \quad (4)$$

Where ω_{ICR} is the hyperparameter used to adjust the effect of the ICR regularization term by adjusting the weights.

We can balance the contribution of the ICR regularization term to the overall loss. This loss function aims to improve the similarity of samples within the same category while driving the model to learn a more discriminative feature representation, which helps optimize speaker recognition performance.

3. Model structure

In this experiment, we introduced the structure of the MFA-conformer model. **Figure 2** outlines the MFA-NConformer model architecture. Fbanks features undergo convolutional subsampling to cut computational costs, then go through Linear and Dropout layers, and are further processed by NConformer blocks for global and local feature extraction. Features from each NConformer block are fused via a Concat layer; an Attention Pooling layer assigns differentiated weights to outputs to extract speaker features, which are finally converted into speaker embedding through batch normalization.

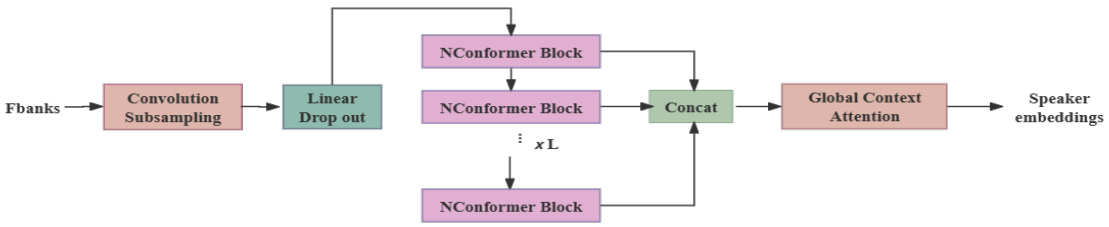

Figure 2. Structure of the MFA-NConformer model.

As shown in **Figure 3**, NConformer is an optimization model based on the conformer model. Adding one-dimensional deep separable convolution and channel attention mechanisms to the feedforward neural network of the conformer block. The resulting NConformer blocks act in the same way as the traditional conformer model. That is, the spatial and temporal features of the input data are learned and feature representations are extracted through multilayer convolution and self-attention mechanisms [14].

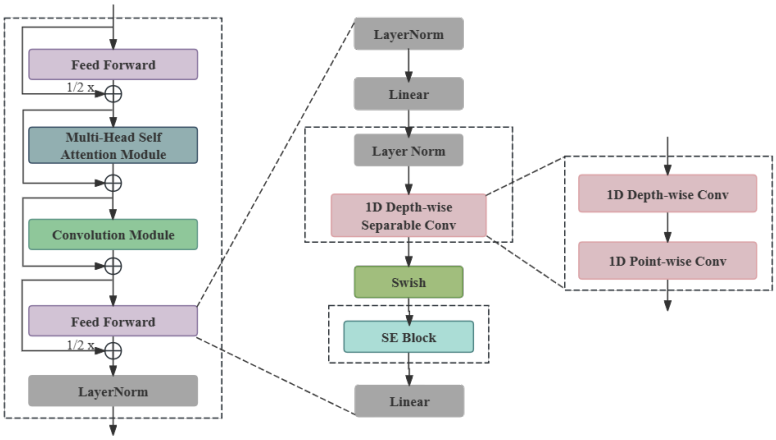

Figure 3. Block diagram of NConformer.

4. Experimental setup and evaluation

We compare the performance of GE2E loss and GE2Eicr loss on the MFA-conformer and MFA-NConformer models, respectively, and contrast 1D convolution with 1D separable convolution in the feedforward layer. Notably, our goal is not SOTA performance but to verify the validity and effectiveness of the "right" margin.

4.1. Tibetan speaker dataset

The experiment's Tibetan corpus was recorded with a professional microphone. Its audio follows a 16 kHz sampling rate, 16-bit mono and WAV format. The corpus contains 47 speakers, each with 500 audio clips, totaling 23,500 clips. The average clip duration is 3 seconds, and the total recording time is over 19 hours. We divide the dataset into a training set of 40 speakers and 20,000 clips, and a test set of 7 speakers and 3,500 clips (**Table 1**).

Table 1. Tibetan speaker dataset

Aspects	Speaker	Utterance/speaker	Utterance
Train	40	500	20000
Test	7	500	3500
Total	47		23500

4.2. Experimental configuration

This experiment used the PyTorch framework, running on CentOS Linux with an Intel Xeon E5-2680 v4 CPU and an 11 GB GeForce RTX 2080 Ti.

4.3. Evaluation indicators

Speaker recognition typically evaluates system performance using three metrics: false acceptance rate (FAR), false rejection rate (FRR), and equal error rate (EER). FAR refers to the proportion of incorrectly accepted data, while FRR denotes the proportion of incorrectly rejected data.

$$FAR = \frac{FP}{FN+TN} \tag{5}$$

$$FRR = \frac{FN}{TP+FN}$$

5. Results and analysis

This experiment compares the MFA-conformer model with its improved variants, models employing the GE2E loss function and the GE2Eicr loss function. Under the same corpus, the equal error rate (EER) was evaluated using different optimizers and pooling layer configurations. **Table 2** shows balanced data, with FBank = 40, batch size = 32, learning rate = 0.001, and a 400-epoch training process. It compares two models' performance under two metric loss functions, with hyperparameter M (fixed at 2, words per speaker). The MFA-NConformer with GE2Eicr loss achieves a 2.6% EER, outperforming the conformer and MFA-conformer. GE2Eicr loss also outperforms GE2E loss, reducing the MFA-NConformer's error rate by 28.96%. We therefore selected this model-loss function combination for subsequent experiments.

Table 2. Comparison of the effects of three models using two loss functions (Adam)

Model	Optimizer	Hyperparameters	EER
MFA-Conformer	GE2E	M=2	4.10
	GE2Eicr		2.80
MFA-NConformer	GE2E		3.61
	GE2Eicr		2.65

According to **Table 3**, when M = 2, the model performs optimally with an EER of 1.90%. However, as M increases, the model's performance declines. A smaller M value allows the loss function to reduce intra-class distances more effectively, resulting in better performance.

Table 3. Effect of M-value on model performance (1D depthwise separable convolution)

Loss	Hyperparameters	EER
GE2Eicr	M=2	1.90
	M=3	2.15
	M=4	2.55

Table 4 compares the results under two different optimizers (Adam and SGD). The results indicate that the SGD optimizer outperforms the Adam optimizer with an EER of 1.77%, making SGD more suitable for this improved model.

Table 4. Effect of M-value on model performance (1D depthwise separable convolution).

Aspects	Optimizer	EER
1D Depthwise Conv	Sgd	1.90
	Adam	3.12
1D Conv	Sgd	1.77
	Adam	2.65

6. Conclusion

This experiment improves Tibetan speaker recognition based on the MFA-conformer model. Introducing 1D separable convolutions (which reduce parameters and computational complexity while ensuring robust feature extraction) and channel attention mechanisms into the feedforward network, integrating multi-conformer block features, and adding an intra-class correlation regularization term to the GE2E loss function, all effectively enhance model performance. These improvements collectively promote the advancement of Tibetan speaker recognition.

Disclosure statement

The author declares no conflict of interest.

References

[1] Boles A, Rad P, 2017, Voice Biometrics: Deep Learning-based Voiceprint Authentication System, In: 2017 12th System of Systems Engineering Conference (SoSE), 1–6.

[2] Chowdhury F, Wang Q, Moreno I, 2018, Attention-based Models for Text-Dependent Speaker Verification, In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5359–5363.

[3] Peddinti V, Povey D, Khudanpur S, 2015, A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts., in Interspeech, 3214–3218.

[4] Sang M, Zhao Y, Liu G, et al., 2023, Improving Transformer-based Networks with Locality for Automatic Speaker Verification, In: ICASSP 2023- 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5.

[5] Cai D, Li M, 2023, Leveraging ASR Pretrained Conformers for Speaker Verification through Transfer Learning and Knowledge Distillation, arXiv, https://doi.org/10.1109/TASLP.2024.3419426

[6] Gulati A, Qin J, Chiu C, et al., 2020, Conformer: Convolution-Augmented Transformer for Speech Recognition, arXiv, https://doi.org/10.48550/arXiv.2005.08100

[7] Zhang Y, Lv Z, Wu H, et al., 2022, Mfa-Conformer: Multi-Scale Feature Aggregation Conformer for Automatic Speaker Verification, arXiv, https://doi.org/10.48550/arXiv.2203.15249

[8] Li L, Wang D, Rozi A, et al., 2017, Cross-Lingual Speaker Verification with Deep Feature Learning, In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 1040–1044.

[9] Wu Y, Liao W, 2021, Toward Text-Independent Cross-Lingual Speaker Recognition using English-Mandarin-Taiwanese Dataset, In: 2020 25th International Conference on Pattern Recognition (ICPR), 8515–8522.

[10] Thienpondt J, Desplanques B, Demuynck K, 2022, Tackling the Score Shift in Cross-Lingual Speaker Verification by Exploiting Language Information, In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7187–7191.

[11] Schuessler A, 2024, Sino-Tibetan in Tibetan and Old Chinese. Language and Linguistics, 80–122.

[12] Mokgonyane T, Sefara T, Manamela M, et al., 2019, The Effects of Data Size on Text-Independent Automatic Speaker Identification System, In: 2019 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), 1–6.

[13] Wan L, Wang Q, Papir A, et al., 2018, Generalized End-to-End Loss for Speaker Verification, In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4879–4883.

[14] Lyu H, Sha N, Qin S, et al., 2019, Advances in Neural Information Processing Systems. Advances in Neural Information Processing Systems, 32(2019).