THE FOSAFER SPEAKER VERIFICATION SYSTEM FOR THE SOURCE SPEAKER TRACING CHALLENGE 2024

Yuxuan Du, Dejun Zhang, Jing Deng, Rong Zheng

Beijing Fosafer Information Technology Co., Ltd., Beijing, China

ABSTRACT

This report presents the system developed by the Fosafer Research team for the voice conversion-based speaker verification task in the Source Speaker Tracing Challenge (SSTC) 2024. To address a variety of known and unknown voice conversion methods, we adopted fine-tuning strategies based on ResNet101 pre-trained model. Furthermore, we evaluated various scoring strategies and training techniques suitable for voice conversion speaker verification task. Our best system achieved an equal error rate (EER) of 18.65% on sixteen test sets, a 9.51% reduction in EER relative to the baseline system. Our final submission secured the second position in SSTC2024.

***Index Terms*—** Speaker verification, voice conversion, fine-tuning strategy

1. INTRODUCTION

The Spoken Language Technology (SLT) conference organized SSTC 2024 to advance the development of speaker verification (SV) techniques, particularly in response to voice conversion (VC) attacks[1]. Over the past few years, several challenges have driven technological advances in the prevention of spoofing in speaker verification systems. In particular, the ASVspoof Challenge and the Audio Deepfake Detection Challenge have played an important role in evaluating and enhancing the ability of automated speaker verification systems to counter forgery attacks, such as speech synthesis, speech transcription, and audio recording playback.

The challenge provided a large-scale, publicly accessible database of converted speech to support research in source speaker verification. The challenge includes 16 types of voice conversion, with the trials for speaker verification subdivided into four categories. The first category is where the two converted voices originate from the same source speaker and the same target speaker. The second category involves two converted speeches from different source speakers but the same target speaker. The third category is where the two converted speeches originate from the same source speaker but have different target speakers. The fourth category involves two converted speeches from different source speakers and different target speakers. During evaluation, the system must determine whether the two converted speech segments in these four scenarios originate from the same source speaker[1]. This paper outlines the various speaker verification systems we utilized for multiple speech conversion methods.

2. SYSTEM DESCRIPTION

2.1. System Training

The training of the model is divided into two phases. The first phase is the initial training phase, where normal speech is used for the training data, aiming for the model to initially learn the general features and obtain better performance in subsequent tasks. The second phase involves fine-tuning, using both officially voice conversion data.

For the initial training phase, VoxCeleb2[2] was used as the training dataset. The ResNet101[3] architecture was used as the backbone network for embedding extractor. To enhance the robustness of the system, two data enhancement techniques were employed. These techniques involved data augmentation using the MUSAN[4] and RIR[2] datasets and applying speed perturbation to triple the number of speakers[5]. The speech features comprised 80-dimensional FBank features with the window size set to 25 ms and the frame shift set to 10 ms. An SGD optimizer was utilized with a momentum of 0.9 and weight decay of 1e-4. The baseline model employed an Angular Additive Margin Softmax (AAM-Softmax) loss[6], with the margin and scaling factors set to 0.2 and 32, respectively. The mean and standard deviation vectors of the frame-level features were computed and concatenated using the Temporal Statistical Pooling (TSTP) function. An exponential decay scheduler was used with an initial learning rate of 0.1 and a final learning rate of 5e-5. In addition, the margin was gradually increased from 0 to 0.2. The first training stage lasted for a total of 150 epochs, with the final 256-dimensional speaker embeddings extracted from the first fully connected layer. For the fine-tuning phase, the model was fine-tuned using the officially provided training set for voice conversion speech. Since voice conversion speech is more difficult to train, speed perturbation was removed to avoid the effect of mismatched test data. The initial learning

[1]https://sstc-challenge.github.io/

[2]https://www.openslr.org/28

rate was set to 0.01, and the final learning rate was set to 5e-5. Considering that too large a margin penalty may negatively impact model performance, the margin was consistently set to 0 in the second stage. The second training stage lasted for a total of 25 epochs. In addition, considering that there are more short-duration speech in the development set, an additional Large Margin Fine-tuning (LMF)[7] training phase was not introduced.

2.2. System Scoring

Cosine similarity served as the scoring criterion. Adaptive Score Normalization (AS-Norm)[8] was applied post-score calculation, with the impostor cohort size set to 200. The cohort was estimated using Librispeech voice conversion training set. Additionally, Quality Measure Functions (QMF) were excluded as they did not enhance performance.

3. RESULTS

The performance of the proposed speaker verification system on the SSTC development and test sets is shown in Table 1. The experimental results show that the baseline system achieved an EER of 19.40% on the development set and 20.61% on the test set. When the fine-tuned model is used to extract speaker embeddings on the development and test sets and AS-Norm is applied, the EER reached 18.39% and 18.64% on the two datasets, representing a relative reduction of 5.21% and 9.51%, respectively, compared to the baseline system.

Table 1. The speaker verification system performs on the SSTC development set and test set.

Index	Dev EER(%)	Test EER(%)
Baseline	19.40	20.61
Our	18.39	18.65

The EER of the baseline system and the proposed systems on different development sets are presented in Figure 1. The experimental results show that the difficulty of the first six development sets is lower than that of the last six. However, despite our proposed system achieving a lower EER on the final development set, the fine-tuning strategy employed by the baseline outperforms our system's performance on the first six development sets. Although the MFA-Conformer pre-trained model used by the baseline had a higher EER on VOX-1 than the ResNet101 pre-trained model we used, it did not seem to affect the fine-tuning performance on the first six development sets. Even though the data tested in DEV-7 and DEV-8 were included in the training, these voice conversion methods are more detrimental to the speaker information and still fail to accurately recognize the source speaker.

4. CONCLUSION

This paper presents the speaker verification system submitted by the Fosafer Research team to SSTC2024. Our final submission secured second place in the competition with an EER of 18.65%. We utilized improved pre-trained models fine-tuned on the voice conversion dataset, which yielded better results on the challenging dataset. Additionally, we employed various scoring strategies to enhance system performance and conducted analyses to uncover interesting insights in the voice conversion speaker verification task. The results indicate significant room for improvement in addressing the impact of partial voice conversion on speaker verification. In the future, we will focus on speaker verification tasks for partially challenging datasets and explore the impact of different systems on various source and target speakers.

5. REFERENCES

[1] Yanzhen Ren, Hongcheng Zhu, Liming Zhai, Zongkun Sun, Rubing Shen, and Lina Wang, "Who is speaking actually? robust and versatile speaker traceability for voice conversion," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8674–8685.

[2] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[5] Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] Xu Xiang, Shuai Wang, Houjun Huang, Yanmin Qian, and Kai Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1652–1656.

[7] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn

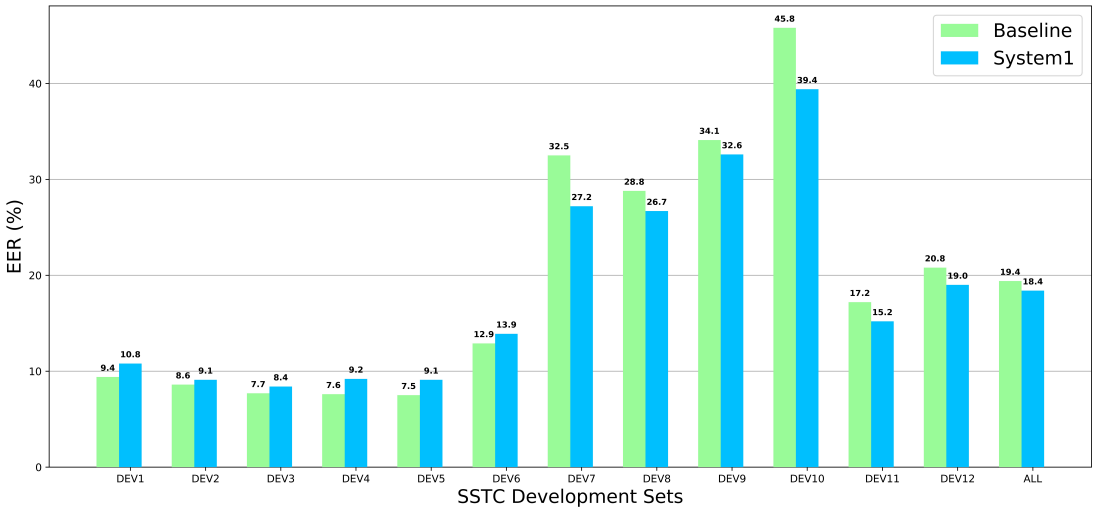

Fig. 1. The EER of the baseline system and the proposed systems on different development sets.

based speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5814–5818.

[8] Sandro Cumani, Salvatore Sarni, et al., "From adaptive score normalization to adaptive data normalization for speaker verification systems," in *Proceedings of the 24th INTERSPEECH Conference*. ISCA, 2023.