THE TIANJIN UNIVERSITY AUDIOVISUAL COGNITIVE COMPUTING TEAM SPEAKER VERIFICATION FOR THE SSTC2024

Hangming Zhang, Zheng Li, Qianyi Bai, Zhichao Deng

College of Intelligence and Computing
Tianjin University
Tianjin, 300354, China

ABSTRACT

In recent years, speaker verification technology has become a common identity verification method, which is widely used in fields like AI assistant, financial security, smart city, criminal investigation. While with the development of artificial intelligence technology like speech conversion, the speaker verification systems are facing more threats than ever before. The Source Speaker Tracing Challenge (SSTC) plays an important role in addressing these problems, providing a platform for communication and innovation of source speaker verification, also providing impetus for the application of new technology in this area. In this challenge, we team *Tianjin University Audiovisual Cognitive Computing* proposed a system which is based on ResNet152 derived from the WeSpeaker framework. In the training process of the model, AAM classifier was used and, 8 data sets were used to simulate different deception scenarios to fine-tune the model, so that the model has ability to recognize and counter complex attacks. The model achieved an ERR of merely 19.3% on the test set of voice conversion speaker verification, ranking the third place in SSTC 2024. In the future, we will continue to explore various ways to further improve the robustness and effective recognition to against more sophisticated deception scenes.

Index Terms— Source speaker verification, speaker verification, the Source Speaker Tracing Challenge (SSTC)

1. INTRODUCTION

In today's digital era, speaker verification[1] has emerged as a crucial biometric authentication technology, extensively utilized in various applications such as mobile devices[2], smart homes[3], smart cities[4], criminal investigation[5] and financial services[6]. These technologies rely heavily on the accuracy and security of voice recognition systems to ensure the integrity of user interactions. However, with the advancement of deep neural networks, the susceptibility of speaker verification systems to spoofing attacks—such as speech synthesis][7], voice conversion, and speech editing—has become an increasing concern. To address these vulnerabilities, various challenges like ASVspoof and Audio Deepfake Detection have been organized, aiming to foster the development of effective countermeasures.

Amidst these developments, the Source Speaker Tracing Challenge (SSTC) 2024 plays a pivotal role. This challenge is specifically designed to enhance the source speaker verification process, particularly against sophisticated voice conversion attacks. By focusing on the identification of the source speaker in manipulated speech signals, SSTC 2024 seeks to push the boundaries of current technologies, promote innovation, and provide a robust, open dataset for community engagement.

This report introduces a novel speaker verification system specifically designed for voice-converted speech. The system has been trained using the official training set provided by the challenge organizers. Notably, without the simulation of additional training sets, our best model achieved an Equal Error Rate (EER) of 19.32%. The structure of this report is organized as follows: Section 2 details our training and fine-tuning strategies, Section 3 presents the experimental results, and Section 4 concludes with a summary of our findings and implications for future research. This system's development and evaluation are part of a broader effort to enhance the reliability and security of speaker verification technologies in the face of evolving spoofing tactics.

2. TRAINING STRATEGY

Our speaker verification system leverages the pretrained ResNet152[8] model as its foundation. This model is sourced from the WeSpeaker framework[9]. The ResNet152 model was trained on the VoxCeleb2 dataset[10]. VoxCeleb2 contains over 1 million utterances from 5994 celebrities, sourced from YouTube videos. The dataset encompasses a wide range of accents, professions, and different backgrounds, making it one of the most diverse and comprehensive datasets available for speaker recognition tasks.

To fine-tune the speaker verification system, we used a specialized dataset created from the LibriSpeech corpus provided by the contest organizers. The converted speech sam-

Table 1. EER of two systems on 12 Dev sets.

System	DEV1	DEV2	DEV3	DEV4	DEV5	DEV6	DEV7	DEV8	DEV9	DEV10	DEV11	DEV12	ALL
S1	12.82	12.03	9.41	11.15	11.41	14.65	31.92	30.78	36.78	42.88	19.16	23.57	21.38
S2	11.78	10.09	8.53	11.17	10.08	14.05	28.22	27.78	33.04	40.82	14.96	19.27	19.15

ples were generated using advanced voice conversion (VC) techniques, which manipulate the vocal characteristics of the source speaker's audio to resemble those of a target speaker while preserving the linguistic content. This process involved selecting three source speech samples for each target speech, simulating scenarios where attacks are carried out by different voice converters. The diversity in conversion techniques ensures that our system is exposed to a wide range of possible spoofing scenarios, enhancing its ability to generalize across different types of voice conversion attacks.

In our research, we employed an advanced fine-tuning strategy to optimize the ResNet152 model for speaker verification tasks, specifically focusing on voice-converted speech. The model was fine-tuned on eight distinct datasets, each representing different voice conversion scenarios to ensure broad exposure and robustness against various spoofing types. The training was conducted over 20 epochs with a warmup phase of one epoch. Spectral features were extracted using an 80-dimensional Fbank. The ResNet152 model was trained with the Additive Angular Margin(AAM) classifier[11], configured with an angular margin m of 0.2 and a scaling factor s of 32, which sharpens the decision boundaries between classes, crucial for improving the discriminative power of the model in distinguishing between different speakers. No dropout was applied, allowing the model to train on all features for a comprehensive learning experience. Our model operated on four GPUs. This fine-tuning approach not only adapted the model to the intricacies of voice-converted speech but also enhanced its generalization capabilities, essential for deploying in real-world scenarios where speaker verification systems must reliably identify and counteract sophisticated spoofing attacks.

3. RESULTS

We have carefully constructed and submitted two systems S1 and S2 to evaluate their performance. The main difference between these two systems lies in a crucial step in the data preprocessing stage, which has a significant impact on the final performance. Firstly, S1 followed the standard processing flow without introducing additional data transformation steps. However, despite S1's outstanding performance in multiple aspects, there is still room for optimization in certain details. To further improve performance, we developed S2. Compared to S1, S2 has added a key step in the data preprocessing stage: calculating the mean vector of all data in the Dev set, and subtracting this mean vector when extracting each embedding. This step aims to eliminate certain biases or noise in the data so that embedding can more accurately reflect the inherent characteristics of the data.

Through testing on the actual development set, we found that system S2 has significantly improved the key indicator of EER, as shown in Table 1. This result fully demonstrates the effectiveness of our adjustments during the data preprocessing stage. After completing the evaluation on the Dev set, we decided to submit S1 and S2 to the Test set for more rigorous validation. The ERR of S1 and S2 on the Test set were 23.20% and 19.32%, respectively. The significant improvement in this achievement once again proves the effectiveness of our work in data preprocessing and model optimization. We believe that with the continuous advancement of technology and in-depth research, our system will be able to achieve even better performance in the future.

4. CONCLUSION

Our team ultimately secured third place in the SSTC, achieving an EER of 19.32% on the test set for voice conversion speaker verification. It is observable that despite fine-tuning the dataset, the experimental results were still not entirely satisfactory. This task remains exceedingly challenging with current technologies, indicating significant room for improvement in handling voice conversion spoofing. Moving forward, we are committed to exploring various training strategies to better understand their impact on voice conversion speaker verification systems. By continually refining our approach, we aim to enhance the robustness and accuracy of our system against sophisticated spoofing techniques.

5. REFERENCES

[1] Craig S. Greenberg, Lisa P. Mason, Seyed Omid Sadjadi, and Douglas A. Reynolds, "Two decades of speaker recognition evaluation at the national institute of standards and technology," *Comput. Speech Lang.*, vol. 60, no. C, mar 2020.

[2] Joao Antônio Chagas Nunes, David Macêdo, and Cleber Zanchettin, "Am-mobilenet1d: A portable model for speaker recognition," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[3] Yudi Dong and Yu-Dong Yao, "Secure mmwave-radar-based speaker verification for iot smart home," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3500–3511, 2020.

[4] Adil E Rajput, Tayeb Brahimi, and Akila Sarirete, "Automatic speaker verification, zigbee and lorawan: Potential threats and vulnerabilities in smart cities," in *Research & Innovation Forum 2019: Technology, Innovation, Education, and their Social Impact 1*. Springer, 2019, pp. 277–285.

[5] Joseph P. Campbell, Wade Shen, William M. Campbell, Reva Schwartz, Jean-Francois Bonastre, and Driss Matrouf, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 95–103, 2009.

[6] Muteb Aljasem, Aun Irtaza, Hafiz Malik, Noushin Saba, Ali Javed, Khalid Mahmood Malik, and Mohammad Meharmohammadi, "Secure automatic speaker verification (sasv) system through sm-altp features and asymmetric bagging," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3524–3537, 2021.

[7] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[8] 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.

[9] 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.

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

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.