SASV Technical Report

Jiahao Zheng, Jinke Ren[†]
FNii-Shenzhen and SSE, CUHK-Shenzhen
Email: jinkeren@cuhk.edu.cn

Abstract

Existing solutions for the Spoof-Aware Speaker Verification (SASV) task are typically trained and tested on clean data collected in studio environments, resulting in suboptimal performance in natural environments. The WildSpoof SASV challenge considers the SpoofCeleb dataset [1] and requires distinguishing among target speech, non-target speech, and spoofed speech in natural environments, posing significant challenges that demand both high detection accuracy and robust interference resistance from detection models. To address this dual challenge, we employ the ECAPA-TDNN model [2] to compare the Mel-spectrograms of source speech with those synthesized by a high-quality vocoder, thereby accentuating the forgery artifacts in speech. Moreover, to mitigate interference from inherent speech noise, we segment the spectrogram into distinct frequency bands and subsequently perform speaker recognition and audio authenticity verification within each band. Experimental results demonstrate that compared to the baseline SKA-TDNN model [1], our proposed model achieves a reduction in the a-DCF value of 0.1512 in real-world scenarios, validating its strong robustness and detection capability for the SASV task.

I. Preliminary

In recent years, deep learning models represented by the ECAPA-TDNN model [2] have achieved remarkable success in the Spoofing-Aware Speaker Verification (SASV) task. However, existing methods often train and test these models in controlled environments such as recording studios, which fail to reflect the characteristics of real-world scenarios. The Wild Text-to-Speech Synthesis (TTS) dataset, exemplified by the SpoofCeleb dataset, simulates attacks that closely mirror reality, can enhance the robustness of systems designed for the SASV task. In response to the WildSpoof Challenge, we utilize the SpoofCeleb dataset to simulate the input for SASV systems in real-world scenarios, thereby enhancing their security and robustness. Notably, the SpoofCeleb dataset comprises 2.5 million speech clips collected under diverse and noisy conditions in realistic environments, including both genuine and spoofed speech, making it highly representative of authentic settings.

II. Methodology

As shown in Fig. 1, the working mechanism of our proposed method consists of the following five steps:

1) Input the source speech into a high-quality speech codec—FACodec [3]—to obtain the reconstructed voice.
2) Calculate the logarithmic Mel frequency spectra of the source speech and the reconstructed speech, denoted by $\boldsymbol{x}_{\text{original}}$ and $\boldsymbol{x}_{\text{reconstruct}}$.
3) Perform a weighted subtraction between $\boldsymbol{x}_{\text{original}}$ and $\boldsymbol{x}_{\text{reconstruct}}$, which is given by

$$\boldsymbol{x}_{\text{residual}} = \boldsymbol{x}_{\text{original}} - c \times \boldsymbol{x}_{\text{reconstruct}}, \tag{1}$$

   where $c \in [0, 1]$ is a pre-determined coefficient.
4) Feed $\boldsymbol{x}_{\text{residual}}$ into the Multi-Band ECAPA-TDNN model to obtain the speaker embedding.
5) Calculate the cosine similarity between the speaker embedding and the enrolled speaker embedding, and determine whether the source speech belongs to the target speaker. Specifically, if the cosine similarity is greater than a threshold $\tau \in [0, 1]$, the source speech belongs to the target speaker; otherwise, it belongs to a nontarget speaker or the speech is spoofed.

[†] Corresponding author

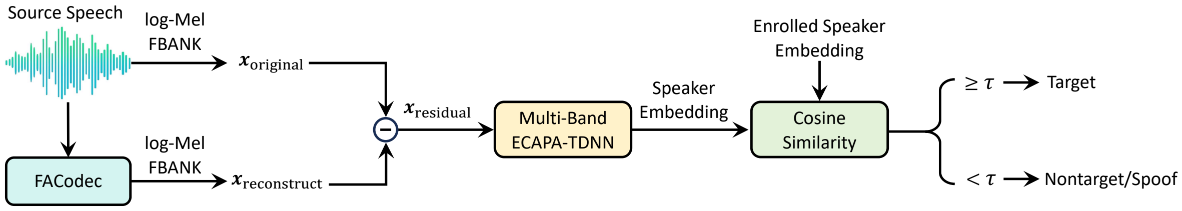
Source Speech
log-Mel FBANK
$\boldsymbol{x}_{\text{original}}$
FACodec
log-Mel FBANK
$\boldsymbol{x}_{\text{reconstruct}}$
−
$\boldsymbol{x}_{\text{residual}}$
Multi-Band ECAPA-TDNN
Speaker Embedding
Enrolled Speaker Embedding
Cosine Similarity
≥ τ
Target
< τ
Nontarget/Spoof

Fig. 1. Working mechanism of the proposed method.

III. Training Algorithm

We use the end-to-end SASV loss function provided by the WildSpoof Challenge for model training. This loss function is composed of two parts: the AAMsoftmax [5] loss and the Angleproto [6] loss. Let $\boldsymbol{v}$, $\boldsymbol{y}$, and $\boldsymbol{\Theta}$ denote the speaker embedding output by the model, the corresponding speech pair label, and the model parameters. Then, the training loss function can be expressed as:

$$L_{\text{SASV}}(\boldsymbol{\Theta}) = \text{AAMsoftmax}(\boldsymbol{v}, \boldsymbol{y}; \boldsymbol{\Theta}) + \text{AngleProto}(\boldsymbol{v}, \boldsymbol{y}; \boldsymbol{\Theta}). \tag{2}$$

Based on this, the training algorithm is shown below.

Algorithm 1: Training Algorithm for the Proposed SASV System

```
Input: Source speech $\boldsymbol{s}_0$, coefficient $c$;
while $L_{\text{SASV}(\boldsymbol{\Theta})}$ not converged do
    $\boldsymbol{x}_{\text{original}} \leftarrow \text{log\_Mel}(\boldsymbol{s}_0)$;
    $\boldsymbol{x}_{\text{reconstruct}} \leftarrow \text{log\_Mel}(\text{Encodec}(\boldsymbol{s}_0))$;
    $\boldsymbol{x}_{\text{residual}} \leftarrow \boldsymbol{x}_{\text{original}} - c \times \boldsymbol{x}_{\text{reconstruct}}$,
    $\boldsymbol{v} \leftarrow \boldsymbol{\Theta}(\boldsymbol{x}_{\text{residual}})$
    Compute $L_{\text{SASV}}(\boldsymbol{\Theta})$ via (2);
    Update $\boldsymbol{\Theta}$;
end
Output: Trained parameters: $\boldsymbol{\Theta}^*$
```

IV. Test Results

We choose SpoofCeleb as the training and validation dataset. We select a-DCF as the evaluation metric. We select the SKA-TDNN model as the benchmark method, which is trained using the official code of the WildSpoof Challenge. The test results are as follows:

TABLE I
Test results of the proposed model and the SKA-TDNN model

Method	a-DCF
SKA-TDNN	0.4183
Ours	0.2671

From the table below, it is observed that compared with the baseline model, our model can achieve lower a-DCF values, achieving better detection performance in real-world environment.

References

[1] Jung, J. W., Wu, Y., Wang, X., Kim, J. H., Maiti, S., Matsunaga, Y., ... Watanabe, S. (2025). SpoofCeleb: Speech deepfake detection and SASV in the wild. IEEE Open Journal of Signal Processing.

[2] Desplanques, B., Thienpondt, J., Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv preprint arXiv:2005.07143.

[3] Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., ... Zhao, S. (2024). Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. arXiv preprint arXiv:2403.03100.

[4] Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., Tagliasacchi, M. (2021). Soundstream: An end-to-end neural audio codec. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 495-507.

[5] Deng, J., Guo, J., Xue, N., Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4690-4699).

[6] Heo, H. S., Lee, B. J., Huh, J., Chung, J. S. (2020). Clova baseline system for the voxceleb speaker recognition challenge 2020. arXiv preprint arXiv:2009.14153.