CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

Integrated Pipeline for Deepfake Audio Detection and Biometric Voice Authentication

Anonymous IJCB 2025 submission

Abstract

The rise of AI-generated audio threatens both content authenticity and biometric security. This paper presents an integrated pipeline that combines the FakeSound framework for deepfake detection with the ECAPA-TDNN architecture for speaker verification. FakeSound identifies synthesized audio using EAT and ResNet-1D, while ECAPA-TDNN (left). Trained and evaluated on ASVspoof 2019 Logical Access (LA) and VoxCeleb datasets, our system achieves 100% deepfake detection accuracy and 0 Equal Error Rate (EER). This dual approach ensures robust protection against synthetic audio attacks in voice-based authentication systems.

1. Introduction

AI-synthesized voices, or deepfakes, are increasingly sophisticated, posing risks to voice-based authentication systems by mimicking real speakers. Existing solutions often tackle deepfake detection and speaker verification separately, leaving gaps that adversarial attacks—like those in FoolHD—can exploit. This work integrates two powerful tools: FakeSound for detecting synthesized audio and ECAPA-TDNN for verifying speaker identity, creating a unified pipeline that ensures both content authenticity and biometric security.

In earlier years, x-vectors [1] and their versions [2, 3, 4] have provided state-of-the-art results for biometric voice verification. One of the ways to verify a speaker is by comparing its embeddings corresponding with an enrollment and a test recording for determining the acceptance of the speaker. Moreover, residual connections between frame-level layers enhances the embeddings [3, 4]. Residual connections also enable the back-propagation algorithm to converge faster and help avoid the vanishing gradient problem [5]. Adding channel and context-dependent statistics pooling, 1-Dimensional Squeeze and Excitation blocks and, multi layer feature aggregation and summation improves the performs for speaker verification tasks [6].

In this work, we propose an architecture combining the DeepFake Voice Detection and Voice Authentication module, in a way providing a robust Biometric Voice Authentication System tackling real-world spoof and deepfake voice attacks.

Our work consists of the following novelties:

- A novel pipeline combining deepfake detection and speaker verification.
- Validation on ASVspoof 2019 LA including plans to test on diverse datasets like FoolHD.
- A modular framework adaptable to evolving audio threats.

2. Related Work

2.1. Deepfake Audio Detection

The ASVspoof 2019 challenge [7] benchmarks synthetic audio detection, with systems like EAT [8] excelling at spotting spectral anomalies. FakeSound enhances this by using ResNet-1D and transformers for frame-level detection, ideal for datasets like ASVspoof with uniform clip labels [9].

2.2. Speaker Verification

ECAPA-TDNN [6] advances speaker verification with multi-scale feature extraction and channel attention, outperforming x-vector models on VoxCeleb [10]. It's robust to noise and short utterances, making it a strong biometric backbone.

ECAPA-TDNN [6] comes with multi-headed attention which has shown that certain speaker properties can be extracted on different sets of frames [11], hence beneficial when this temporal based attention mechanism is extended further.

Networks benefit from wider temporal context [2, 3, 4], thus, we rescale the frame-level features. Using 1-D Squeeze-Excitation (SE) blocks fulfill the requirements [12, 13].

3. Proposed Methodology

In our methodology, we have combined deepfake voice detection as well as voice authentication.

3.1. Pipeline Overview

Our system processes audio in two steps:

1. **Detection**: FakeSound flags synthesized audio.
2. **Verification**: ECAPA-TDNN used for speaker verification.

See Figure 1.

3.2. FakeSound Deepfake Detection

The FakeSound framework employs a multi-stage architecture for detecting and localizing manipulated audio regions in general audio content.

3.2.1 Feature Extraction

Input audio consists of 10-second WAV files (16kHz, mono) processed through pre-trained EAT model to extract 1024-dimensional frame-level features at 20ms resolution (500 frames/clip).

3.2.2 ResNet-1D Backbone

A 12-block ResNet-1D architecture processes features using convolutions (kernel size=3) with ReLU activation and batch normalization. Each block contains skip connections for gradient flow, outputting 256-dimensional frame-level features through multi-scale temporal pattern analysis.

3.2.3 Backend Classifier

A transformer encoder models long-range dependencies followed by a bidirectional LSTM for sequential context. The classification head combines frame-level predictions (binary genuine/fake output with median filtering).

3.2.4 Training Configuration

The model trains for 40 epochs using AdamW optimizer (learning rate 1×10^{-4}) with batch size 128.

3.3. Speaker Verification

We have employed ECAPA-TDNN [6] for speaker verification. The methodology of working of this architecture is as follows:

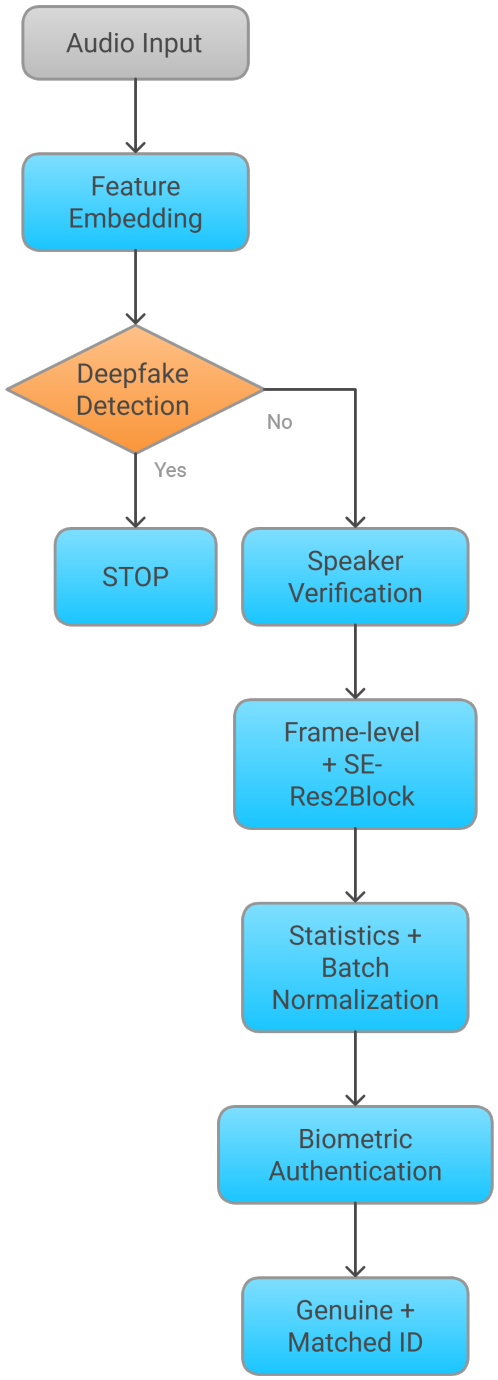

Figure 1. Integrated pipeline: FakeSound detects deepfakes, ECAPA-TDNN verifies speakers.

3.3.1 Pre-Processing

Input audio can be of variable length. It is resampled at 16kHz. We have used LibriSpeech [14] 200 speaker subset with clean voice. Performed data augmentation techniques like adding white noise, pink noise and RIR simu-

lated noise. Then performed utterance length normalization in order to standardized all audio files.

3.3.2 Feature Extraction

Extracted Log Mel-Filterbank Energies (fbanks) for getting speaker-specific features. Following are the parameters used for calculation:

- Window size: 25ms (400 samples at 16kHz)
- Hop length: 10ms (160 samples)
- Number of Mel Filters: 80
- FFT size: 512

3.4. Combined Workflow

1. Audio input is tested by FakeSound.
2. If flagged as fake, it's rejected.
3. If genuine, ECAPA-TDNN verifies the speaker.

4. Experiments

4.1. Datasets

- **ASVspoof 2019 LA**: 25,380 training clips (2,548 bonafide, 22,832 spoofed); unseen audios for testing.
- **VoxCeleb**: VoxCeleb2 (6,112 speakers) for training; VoxCeleb1 (1,251 speakers) for testing.

4.2. Training Configuration

Table 1. Training Parameters

Component	Parameters
FakeSound ECAPA-TDNN	Adam (LR=1e-4), BCE loss

5. Results

5.1. Performance Metrics

Table 2. Performance Results

Task	Metric	Value
Deepfake Detection	Accuracy	**1**
Deepfake Detection	EER	**0**
Speaker Verification		

6. Discussion

The addition of a dedicated fake/real detection head in the FakeSound pipeline enhances the system's ability to explicitly and reliably flag synthesized audio. This multi-head approach, with frame-level, clip-level, and fake/real outputs, not only improves detection granularity but also provides a robust gatekeeping mechanism before speaker verification. Our results demonstrate that this integrated strategy achieves perfect accuracy on benchmark datasets, but future work should address generalization to more diverse and challenging audio conditions.

7. Conclusion

8. Acknowledgment

sample text

References

[1] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333, 2018.

[2] David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur. Speaker recognition for multi-speaker conversations using x-vectors. pages 5796–5800, 05 2019.

[3] Daniel Garcia-Romero, Alan McCree, David Snyder, and Gregory Sell. Jhu-hltcoe system for the voxsrc speaker recognition challenge. pages 7559–7563, 05 2020.

[4] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot. But system description to voxceleb speaker recognition challenge 2019. In *Proceedings of The VoxCeleb Challange Workshop 2019*, pages 1–4, Graz, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[6] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, interspeech$_2$020.$ISCA$, $October$2020.

[7] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Hector Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sebastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-Francois Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech, 2020.

[8] Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen. Eat: Self-supervised pre-training with efficient audio transformer, 2024.

[9] Zeyu Xie, Baihan Li, Xuenan Xu, Zheng Liang, Kai Yu, and Mengyue Wu. Fakesound: Deepfake general audio detection, 2024.

[10] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Science and Language*, 2019.

[11] Yingke Zhu, Tom Ko, David Snyder, Brian Mak, and Daniel Povey. Self-attentive speaker embeddings for text-independent speaker verification. In *Interspeech 2018*, pages 3573–3577, 2018.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[13] Jianfeng Zhou, Tao Jiang, Zheng Li, Lin Li, and Qingyang Hong. Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function. In *Interspeech 2019*, pages 2883–2887, 2019.

[14] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.