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 [2–4] 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.

Architecture of ECAPA-TDNN contains 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 [2–4], 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

ECAPA-TDNN [6] is used for speaker verification. The working procedure of the architecture is in the following subsections.

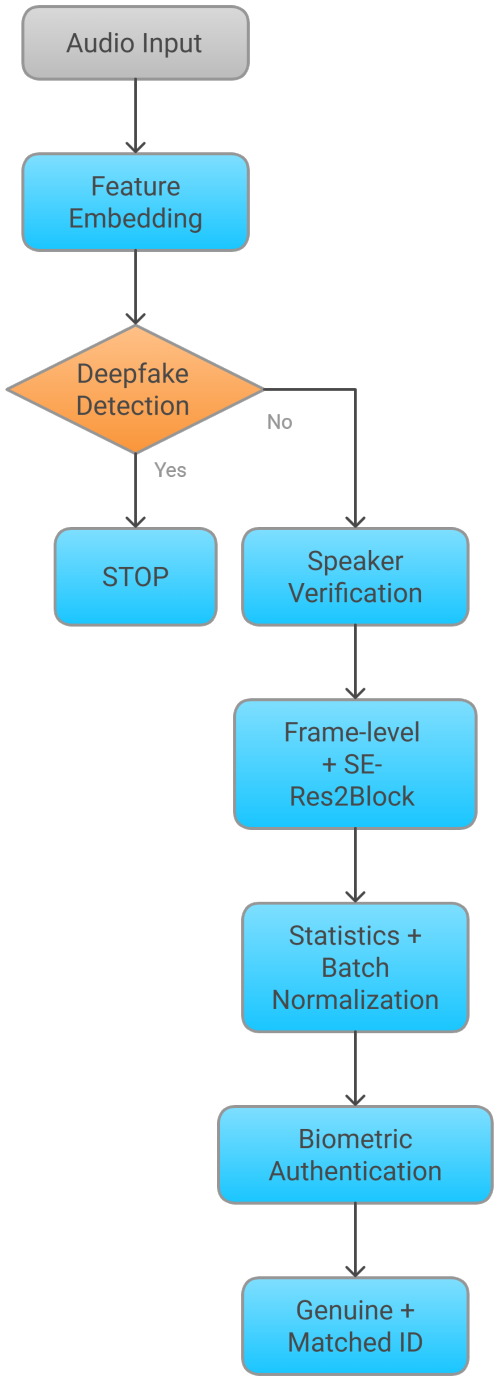

Figure 1. Integrated pipeline: FakeSound detects deepfakes, ECAPA-TDNN verifies speakers.

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-

Table 1. ECAPA-TDNN Architecture

Layer	Input	Output	Role
Input	$[B, 1, 80, \sim 300]$	$[B, 1, 80, \sim 300]$	Log Mel-filterbanks (80 bins, $\sim$3 s).
Conv1D	$[B, 1, 80, \sim 300]$	$[B, 1024, 80, \sim 300]$	1D conv (kernel=5, $C = 1024$).
SE-Res2Net (x3)	$[B, 1024, 80, \sim 300]$	$[B, 1024, 80, \sim 300]$	Multi-scale conv (dilation=2,3,4, scale=8), Squeeze-Excitation.
Concat	$[B, 1024, \sim 300]$ (x3)	$[B, 3072, \sim 300]$	Combines SE-Res2Net outputs.
Conv1D (Agg.)	$[B, 3072, \sim 300]$	$[B, 1024, \sim 300]$	1D conv (kernel=1).
ASP	$[B, 1024, \sim 300]$	$[B, 2048]$	Attention-weighted mean/std (2×1024).
FC	$[B, 2048]$	$[B, 192]$	192D embedding for verification.

Table 2. Training hyper-parameters of ECAPA-TDNN

Parameter	Value
Epochs	50 ($\sim$24 min)
Batch Size	32
Optimizer	Adam
Loss	AAM-softmax ($m = 0.2$, $s = 30$)
Dropout	0.3
Weight Decay	1e-5

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

3.3.3 Model Architecture

We have done some little tweaks in order to accomodate architecture with LibriSpeech dataset. Batch-size (B) is 32 as shown in table 1.

3.3.4 Training

Trained model twiced, one for 50 epochs where early stopping triggered at epoch 15, and another one for 50 epochs without early stopping module.

3.4. Combined Workflow

Working of combination of DeepFake detector and voice verifier can be described as:

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

The flow-chart mentioned in figure 1 represents the same.

4. Experiments

4.1. Datasets

- **ASVspoof 2019 LA**: 25,380 training clips (2,548 bonafide, 22,832 spoofed); unseen audios for testing.
- **LibriSpeech**:
  - Used *train-clean-360*, a subset of LibriSpeech [14], contains 251 speakers.
  - We took 200 speakers for training, of whom male and female speakers were equally divided.
  - 50 utterances per speaker were used, totaling 2 to 3 minutes of speech per speaker.
  - Train-test split ratio - 4:1.

4.2. Training Configuration

The model hyper-parameters of the DeepFake Detector module is in table 3, and those of the Speech Verification module are in table 2.

Table 3. Training hyper-parameters of FakeSound

Parameter	Value
Epochs	40
Batch Size	128
Optimizer	AdamW
Learning Rate	0.0001
Loss	BCE

5. Results

The results obtained on DeepFake detection module is shown in table 4. And, the results obtained on Voice verification module is shown in table 5.

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

Table 4. DeepFake Detection Performance

Dataset	Accuracy (%)	EER (%)
Evaluation	100.0000	0.0000

Table 5. ECAPA-TDNN Performance

Training	Accuracy (%)	EER (%)
Early Stopping at epoch 15	92.9750	1.0500
50 Epochs	92.3750	0.2000

future work should address generalization to more diverse and challenging audio conditions.

7. Conclusion

References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[2] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5796–5800, 2019.

[3] D. Garcia-Romero, A. McCree, D. Snyder, and G. Sell, "Jhu-hltcoe system for the voxsrc speaker recognition challenge," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7559–7563, 2020.

[4] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," in *Proceedings of the VoxCeleb Challenge Workshop 2019*, (Graz), pp. 1–4, 2019.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[6] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn-based speaker verification," in *Interspeech 2020*, p. 2650, 2020.

[7] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "Asvspoof 2019: A large-scale public database of synthesized, converted, and replayed speech," 2020.

[8] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "Eat: Self-supervised pre-training with efficient audio transformer," 2024.

[9] Z. Xie, B. Li, X. Xu, Z. Liang, K. Yu, and M. Wu, "Fakesound: Deepfake general audio detection," 2024.

[10] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[11] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification," in *Interspeech 2018*, pp. 3573–3577, 2018.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

[13] J. Zhou, T. Jiang, Z. Li, L. Li, and Q. Hong, "Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function," in *Interspeech 2019*, pp. 2883–2887, 2019.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.