HUBERT-DERIVED SSL FEATURES AND ECAPA-TDNN MATCHING FOR ROBUST AUDIO DEEPFAKE DETECTION

Gul Tahaoglu

Karadeniz Technical University
Department of Computer Engineering
Trabon, Turkey

ABSTRACT

The rapid advancement and growing accessibility of deepfake audio technologies have prompted substantial concerns, particularly in domains such as politics and media, regarding the reliability of distinguishing between authentic and manipulated audio recordings. This study proposes a robust deepfake audio detection framework combining self-supervised learning (SSL) features extracted using HuBERT models and a powerful ECAPA-TDNN classifier enhanced with One-Class Softmax (OC-Softmax). Three HuBERT variants—Base, Large, and XLarge—were assessed, along with various fusion strategies. Experiments were conducted on the ASVspoof 2019 LA dataset and demonstrated that the proposed system significantly outperforms existing state-of-the-art approaches. The best configuration, based on score level fusion of HuBERT-Large and HuBERT-XLarge with ECAPA-TDNN, achieved an EER of 0.20% and a minimum t-DCF of 0.006. Available at: `https://github.com/gultahaoglu/Hubertderiveredfeatures_deepfakeaudiodetection`

***Index Terms*—** deepfake audio, audio spoofing detection, HuBERT-based features, ECAPA-TDNN

1. INTRODUCTION

The development of neural speech synthesis and voice conversion technologies has reached a stage where artificial voices duplicate human speech so accurately that they become indistinguishable from real voices. The fast advancement of technology has revealed major weaknesses in Automatic Speaker Verification (ASV) systems which create substantial cybersecurity threats. The development of effective detection strategies for deepfake audio has become increasingly critical because of the growing need to protect against fake speeches and fraudulent phone calls and tampered audio content. The generation of deepfake audio occurs through three primary methods which include Text-to-Speech (TTS) and Voice Conversion (VC) and replay-based attacks[**?**].

The process of generating artificial speech from written text operates under two names: Text-to-Speech (TTS) and Speech Synthesis (SS). Advanced deep learning models enable TTS systems to create realistic audio output that mimics specific voices even when original recordings are unavailable.

Voice Conversion (VC) differs from the other method because it transforms source speaker vocal attributes to match target speaker characteristics without altering the original linguistic content. Neural architectures separate speaker identity from spoken content to enable flexible voice adaptation during this process.

Replay attacks function as basic yet dangerous methods of deception. An attacker records authentic speech from a valid user before using it to deceive speaker verification systems, which could result in unauthorized system access [**?**].

In recent years, research dedicated to the detection of deepfake audio has notably increased. Most of these studies adopt either a conventional two-stage architecture—comprising a front-end feature extractor followed by a back-end classifier—or an end-to-end framework, where a single model is trained to perform both feature extraction and classification directly from raw audio waveforms. Generally, hand-crafted spectral features and deep learning-based features trained from scratch have been employed during the feature extraction stage. However, with the growing effectiveness of self-supervised learning (SSL)-based representations, such features are now being increasingly adopted in this domain. Although training self-supervised speech models can be computationally expensive, several pre-trained models—such as wave2vec 2.0 [**?**], WavLM [**?**], and HuBERT [**?**]—are readily available. HuBERT comes in multiple variants, including HuBERT-Base, HuBERT-Large, and HuBERT-XLarge. While HuBERT-Base serves as the standard version, HuBERT-Large and HuBERT-XLarge offer increased model capacity and improved performance. The latter utilizes significantly more parameters and training data to enhance representation quality for downstream speech tasks.

In this paper, the proposed systems employ HuBERT-derived features and their various fusion versions, which are then classified using ECAPA-TDNN with OC Softmax.

Anonymous.

979-8-3503-2411-2/25/$31.00 ©2025 IEEE

The rest of the paper is structured as follows: The literature overview is given in the next section. Then, the details of the proposed audio spoofing detection systems are given in the Methodology section. The Experimental Results section illustrates the presentation and analysis of the experimental methodology and results. Finally, the study is concluded in the Conclusion section.

2. RELATED WORKS

Deepfake audio detection systems are generally structured with two primary modules: a front-end feature extractor and a back-end classifier. However, recent advancements have led to the adoption of end-to-end architectures that integrate both stages within a single unified model.

A wide range of handcrafted front-end features employed in deepfake audio detection are predominantly derived from short-term spectral analysis. Commonly used examples include Mel-Frequency Cepstral Coefficients (MFCC) [?], Inverted MFCC (IMFCC)[?], Linear Frequency Cepstral Coefficients (LFCC)[?], Short-Time Fourier Transform (STFT)[?]. These features rely on short, fixed-length analysis windows and are effective in capturing local spectral patterns. To address the limitations of short-term analysis, particularly in modeling long-range temporal dependencies and enhancing frequency resolution, long-term windowing techniques have been introduced. Among these, Constant-Q Cepstral Coefficients (CQCC) have shown promising results across various studies [?]. In addition to cepstral features, spectrogram-based representations have also gained traction for their ability to provide a time-frequency perspective of audio signals. Notably, Mel-spectrograms (Mel-Spec)[?] and Constant-Q Transform spectrograms (CQT-Spec)[?] have been utilized as rich input features for deep learning models. More recently, a shift toward feature extraction via pre-trained self-supervised learning (SSL) models has been observed. Notable models in this category include wav2vec 2.0 [?], WavLM [?], HuBERT [?], and the Whisper encoder [?], which offer robust, context-aware representations learned from large-scale unlabeled speech data.

A variety of approaches have been explored for the classification stage in deepfake audio detection, including both classical machine learning techniques and modern deep learning frameworks. Traditional models such as Support Vector Machines (SVMs) and Gaussian Mixture Models (GMMs) have been extensively applied, owing to their capacity to capture the statistical properties of audio representations [?, ?, ?]. On the other hand, deep learning models—particularly Convolutional Neural Networks (CNNs) and Residual Networks (ResNets)—have demonstrated strong performance in learning hierarchical feature representations directly from input data.

3. METHODOLOGY

In this study, various versions of HuBERT, a Transformer-based model pretrained using the self-supervised learning (SSL) paradigm, were analyzed as feature extractors from audio signals. The extracted representations were then evaluated using an ECAPA-TDNN-based classifier in conjunction with the One-Class Softmax (OC-Softmax) loss function. This framework enabled a comprehensive analysis for the detection of deepfake audio, leveraging the synergy between advanced SSL-based feature extraction and robust speaker verification techniques.

The overall architecture of the proposed method is illustrated in Figure 1. It consists of two main components: (1) the Hubert-derivered SSL feature extraction, as frontend module (2) ECAPA-TDNN-based classification as a backend classifier. The details of each component are provided in the following subsections.

3.1. Feature extractors:Hubert-Derivered SSL Features

The HuBERT (Hidden-Unit BERT) is an SSL model pretrained on the LibriSpeech-960 dataset, consisting of 960 hours of 16kHz speech [?]. It addresses speech representation challenges by using offline clustering to create aligned target labels for a BERT-like prediction loss. It focuses on masked regions to learn combined acoustic and language models. Although all three HuBERT variants employ the same convolutional feature encoder and masked-unit BERT objective, progressively widening and deepening the Transformer backbone produces substantial gains in representational capacity and automatic speech recognition (ASR) accuracy. HuBERT-BASE couples a 12-layer, 768-dimensional encoder with roughly 95 million parameters, providing a lightweight model that already surpasses traditional supervised baselines. Doubling the depth to 24 layers and enlarging the hidden size to 1024 dimensions raises the parameter budget to about 317 million in HuBERT-LARGE, which consistently lowers word-error rate (WER) across LibriSpeech splits while still being deployable on high-end GPUs. The HuBERT-XLARGE configuration extends the same design philosophy further, employing 48 layers with 1280-dimensional representation. The summary of the model architectures are given in Table 1.

Layer-wise analysis of HuBERT-based features indicates that intermediate blocks (layers 6–9) furnish the most discriminative evidence for synthetic artefacts, with the absolute minimum equal-error rate (EER) achieved at layer 8 [?]. The 20 ms-spaced frame-level vectors emitted at this depth retain the subtle spectral-envelope distortions, harmonicity cues and phase discontinuities introduced by neural vocoders, while avoiding the higher-order lexical abstractions that might obscure such artefacts. In line with these observations, the present work extracts the 768-dimensional hidden states pro-

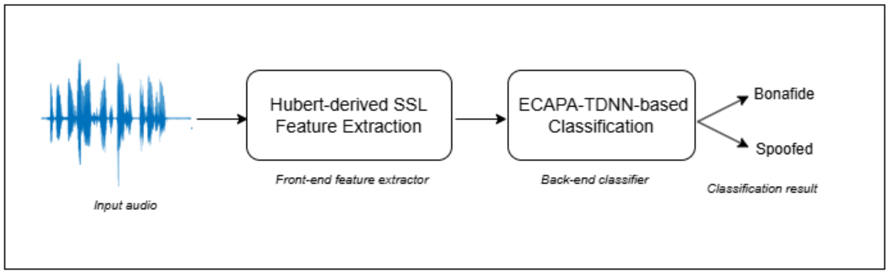
Hubert-derived SSL Feature Extraction
ECAPA-TDNN-based Classification
Bonafide
Spoofed
Input audio
Front-end feature extractor
Back-end classifier
Classification result

Fig. 1. The main framework of proposed audio spoof detection systems

duced by HuBERT's eighth transformer block [?].

Table 1. Model architecture summary for HuBERT-Base, HuBERT-Large, and HuBERT-XLarge

	Component	**BASE**	**LARGE**	**X-LARGE**
CNN Encoder	strides	5, 2, 2, 2, 2, 2, 2		
	kernel width	10, 3, 3, 3, 3, 3, 2		
	channel	512		
Transformer	layer	12	24	48
	embedding dim.	768	1024	1280
	inner FFN dim.	3072	4096	5120
	layerdrop prob	0.05	0	0
	attention heads	8	16	16
Projection	dim.	256	768	1024
	Num. of Params	95M	317M	964M

3.2. Backend Classifier:ECAPA-TDNN

Time Delay Neural Network (TDNN) is a successful x-vector architecture that, together with a statistics pooling layer, transforms variable-duration speech segments into fixed-dimensional embeddings capturing speaker-specific characteristics. The ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation in TDNN), introduced as an enhancement to TDNN, integrates modern architectural innovations to improve the extraction of speaker embeddings from speech signals [?]. The architecture incorporates several key innovations:

- SE-Res2Blocks: These modules combine Res2Net structures with Squeeze-and-Excitation mechanisms to model channel interdependencies and capture multi-scale features, thereby enhancing the network's representational capacity.

- Multi-layer Feature Aggregation: By aggregating features from different layers, the model is able to leverage both low- and high-level acoustic information, resulting in more robust speaker representations.

- Channel-dependent Attentive Statistics Pooling: This pooling mechanism enables the model to focus on temporally informative frames for each channel, improving the quality of the final embedding.

In this paper, OC-Softmax was used as a loss function [?]. The objective of this loss function is to simultaneously encourage compactness among bona fide speech representations and enforce separation from spoofing attacks by employing two distinct margins. This approach is formalized through the One-Class Softmax (OCSoftmax) function, as defined in Equation (1).

$$L_{OCS} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{\alpha(m_{y_i}-\mathbf{w}_0\cdot\mathbf{x}_i)(-1)^{y_i}}\right) \qquad (1)$$

where α denotes a scaling factor, $\mathbf{w}_0$ represents the weight vector, and N is the number of samples in a mini-batch. The vectors $\mathbf{x}_i$ and $\mathbf{y}_i$ denote the input and corresponding embedding vectors, respectively. Two distinct margins, m_0 and m_1 (with $m_0, m_1 \in [-1, 1]$ and $m_0 > m_1$), are employed to characterize bona fide speech and spoofing attacks, respectively.

3.3. Fusion strategies

To quantify how complementary information from different self-supervised front-ends can be exploited, we explored three progressively later fusion stages. Throughout the paper we denote the single-stream reference systems as **S1**(HuBERT-Base), **S2**(HuBERT-Large) and **S3**(HuBERT-XLarge). All three back-ends share an identical ECAPA-TDNN encoder trained with the OC-Softmax loss.

(i) Feature-level fusion

The first strategy concatenates frame-wise features *before* they enter the encoder. Specifically, given two streams $\mathbf{X}^{(a)} \in \mathbb{R}^{C_a \times T}$ and $\mathbf{X}^{(b)} \in \mathbb{R}^{C_b \times T}$, we form $\mathbf{X} = [\mathbf{X}^{(a)} \parallel \mathbf{X}^{(b)}] \in \mathbb{R}^{(C_a+C_b)\times T}$. The fused tensor is processed by a single ECAPA-TDNN, yielding a 256-dimensional utterance embedding and a two-class soft decision. We evaluated **S5**: HuBERT-Large + HuBERT-X-Large As evidenced by the experimental results reported in Table 2, the joint use of

HuBERT-Large and HuBERT-XLarge features yields a pronounced improvement in system performance. Consequently, all subsequent fusion strategies concentrate exclusively on these two feature sets and are implemented as binary fusion schemes.

(ii) Embedding-level fusion

In the second strategy each stream is first encoded *independently* by its own ECAPA-TDNN branch. Let $\mathcal{E}^{(k)}(\cdot)$ denote the branch for stream k; it outputs a normalised embedding $\mathbf{e}^{(k)} \in \mathbb{R}^d$. The embeddings are concatenated and ℓ_2-normalised, $\mathbf{e} = \operatorname{norm}\big[\mathbf{e}^{(1)} \parallel \mathbf{e}^{(2)}\big] \in \mathbb{R}^{2d}$, and classified by a linear layer.

(iii) Score-level fusion

Finally, we combined the decisions of multiple systems by averaging their spoofness scores. Given K calibrated subsystems with scores $s_k(u)$ for utterance u, the fused score is $\bar{s}(u) = \frac{1}{K}\sum_{k=1}^{K} s_k(u)$. In our experiments we fused the two strongest individual models (HuBERT-Large and HuBERT-XLarge) at both the embedding and the score level. The simple arithmetic mean delivered the *best* overall performance, highlighting the robustness of late fusion when constituent systems display diverse error patterns.

4. EXPERIMENTAL RESULTS

This section outlines the experimental setup used to assess the performance of the proposed deepfake audio detection method. In addition, it presents a comparative evaluation against current state-of-the-art techniques documented in the literature.

Each model was implemented using the PyTorch framework. The training procedure employed the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 32. The network was trained for 100 epochs, and the model checkpoint corresponding to the lowest validation EER was selected for evaluation. All experiments were conducted on a laptop with an Intel i7 processor, 64 GB of RAM, and a GeForce RTX 3060 GPU.

4.1. Dataset

The performance of the systems was trained and tested on the ASVspoof 2019 Logical Access (LA) dataset [**?**]. This dataset is central to evaluating the robustness of speaker verification systems against various forms of synthetic and manipulated audio.

The ASVspoof 2019 LA dataset comprises a large diversity of spoofing techniques categorized into three broad categories: Text-to-Speech (TTS), Voice Conversion (VC), and hybrid techniques. TTS-based attacks generate speech from text inputs using deep learning models and neural vocoders to generate highly intelligible and natural-sounding audio. VC attacks produce real speech from the voice of one speaker and transform it into another by passing spectral features through statistical models or deep neural networks. The corpus consists of real recordings of both male and female speakers, which are reference materials used for generating VC samples. Hybrid methods combine TTS and VC methods to generate more advanced and realistic spoofing attacks, often implementing higher-level neural network architectures and signal processing pipelines.

4.2. Evaluation Metrics

To quantify the efficiency of the provided system, two traditional measures of system assessment are utilized: Equal Error Rate (EER) and minimum tandem Detection Cost Function (min-tDCF) [**?**]. They are extensively used in speaker verification and spoofing detection tasks in an effort to express the quantitative impression of system performance.

Both these measures collectively offer an end-to-end evaluation system: Smaller values of EER mean fewer verification errors in terms of balancing acceptance and rejection. Smaller values of min-tDCF indicate the resilience of the system in real-world, integrated settings, where false alarms and missed detections are unacceptable. This two-metric evaluation ensures the recommended system is not just accurate in a lab environment but also practical for actual use, wherein security and reliability matter most.

4.3. Evaluation Results

This section presents a detailed performance evaluation of individual and fused systems, supported by metric results on the dataset.

Table 2 presents each system's EER(%) and min t-DCF scores on the eval set. Among the individual models, S2 (HuBERT-Large+ECAPA-TDNN) delivers the best stand-alone result with an EER of 1.33% and a min t-DCF of 0.035. Fusion further boosts performance: the score-level fusion of S2 and S3 (S6) achieves the overall lowest EER of 1.21 % while matching the best min t-DCF of 0.034. The feature-level fusion of the same pair (S4) follows closely with an EER of 1.23 % and an identical min t-DCF of 0.034. These outcomes highlight that fusion, especially at the score level between strong individual models—consistently enhances anti-spoofing robustness.

Table 2. Test results of each system on the eval set of ASVspoof 2019 LA

ID	Systems	EER (%)	min t-DCF
S1	Hubert-Base +ECAPA-TDNN	3.09	0.099
S2	Hubert-Large +ECAPA-TDNN	0.25	0.007
S3	Hubert-XLarge+ ECAPA-TDNN	0.50	0.012
S4	Feature Level Fusion (S2+S3)	0.99	0.028
S5	Embedding Level Fusion (S2+S3)	0.97	0.027
S6	Score Level Fusion (S2+S3)	0.20	0.006

Table 3 gives the comparative performance results of various state-of-the-art methods evaluated on the ASVspoof 2019 LA dataset, with performance measured in terms of EER and min t-DCF values. The proposed score-level fusion of HuBERT-Large and HuBERT-XLarge-based approach outperforms all compared methods, delivering a lower EER of 0.20% and min t-DCF of 0.006. These results demonstrate the effectiveness and reliability of the final fused system in detecting spoofed audio on the LA dataset.

Table 3. State-of-the-art methods and their performance

State-of-the-art methods	EER (%)	min t-DCF
Zhang et. al. (2021) [?]	2.19	0.05
Li et. al. (2021) [?]	1.78	0.05
Hak et. al. (2021) [?]	4.62	0.12
Hua et. al. (2021) [?]	1.64	0.04
Xue et. al. (2023) [?]	2.73	0.07
Xue et al. (2023) [?]	2.82	0.07
Zaman et. al. (2024) [?]	3.22	-
Wang et al. (2024) [?]	1.94	-
Chaiwongyen et al. (2024) [?]	15.97	-
Mirza et. al. (2024)[?]	6.67	0.21
Final system	**0.20**	**0.006**

5. CONCLUSION

This work presents a comprehensive deepfake audio detection framework leveraging self-supervised HuBERT-derived representations in conjunction with the ECAPA-TDNN classifier trained using OC-Softmax loss. The experiments on the ASVspoof 2019 LA dataset demonstrated that the combination of SSL-based features and the ECAPA-TDNN architecture can substantially improve detection accuracy, especially when employing larger HuBERT variants and fusion strategies. Notably, the final system outperformed existing benchmarks, achieving a minimum EER of 0.20% and a min-tDCF of 0.006, confirming its robustness and generalizability.

6. REFERENCES