An Efficient Approach for Audio Deepfake Detection

Quan Trong The

Lab Blockchain, Faculty Information Security, Posts and Telecommunications Institute of Technology (PTIT), Hanoi, Vietnam

Abstract

Nowadays, with the rapid development of multimedia technology for meeting the requirements of humans, deep learning techniques are used for generating, creating synthetic media. This raises significant challenging problems in controlling audio-based content with deepfake sound. Deepfake makes it harder to identify the original sound source and authentication process in almost all speech applications, such as, voice-controlled devices, teleconference systems, biometric equipment. Much research has attempted to recognize, classify audio deepfake by utilizing the AVSpoof and numerous different machine learning (ML) algorithms with perspective numerical simulations. In this paper, the author proposed applying an optimized Mel-frequency cepstral coefficients (MFCCs) to improve the robustness of deepfake sound detection system, which based on ML in the terms of measured performance metrics, such as accuracy, precision, recall and F1-score. The numerical results have confirmed the effectiveness of the described method in comparison to traditional approaches. The author's suggested technique offers a promising framework for identifying, detecting and classifying deepfake voices to prevent the spread of misinformation of digital media.

Keywords

Audio deepfake, mel-frequency cepstral coefficients, detection, deep learning, voice

1. Introduction

To detect audio deepfake, which is synthesized, edited or generated by using AI, deep-learning is an essential requirement of almost human-interactive interface; for instance, call centers, banking and customer service. For detecting, one must first understand the method, algorithm and technique of generation. Audio deepfake can be categorized into: replay attack, speech synthesis and voice conversion. Replay attacks repeatedly playing back the recording intended victim. As of now, deep convolutional networks [3] are applied to detect in the form of far field detection and copy-and-paste detection [4-5]. The described method was verified with ASVspoof2017 database [6] and has promising results in the term of Equal Error Rate (EER). Speech synthesis (SS) is recreating human voice digitally by implementing computer software and hardware. AI personal assistants and text reading are just two common popular applications of speech synthesis. Beside, speech synthesis can mimic various voices and dialects. To synthesize 1,000 sentences per second, Lyrebird utilizes deep learning models to build the system, creating speech corpora. Char2Wav, PixelCNN, WaveNet are prospective frameworks for speech synthesis. For audio data, [7-8] applies GAN-based generative models to synthesize speech. These above approaches are based on a fully convolution feed-forward network to operate the Mel spectrograms, which comprises 117.985 audio segments of 16-bit Pulse Code Modulation.

Deep Fakes are increasingly detrimental to various business applications, Authenticity, social security and privacy. However, in recent years, deepfake in video detection has attracted more scholars and obtained greater accuracy. Therefore, audio spoofing and calls from malicious sources are generated through deep fakes become a crucial problem, due to deepfake audio detection being less studied than image and video - based approaches. In this article, the author presented direct research utilizing optimized MFCCs features for detecting and classifying based on machine learning Random Forest, SVM algorithms. The numerical simulation compares the result and analyzes the baseline models.

ProfIT AI'25: 5th International Workshop of IT-professionals on Artificial Intelligence, October 15–17, 2025, Liverpool, UK

✉ theqt@ptit.edu.vn (Q. T. The)
iD 0000-0002-2456-9598 (Q. T. The)

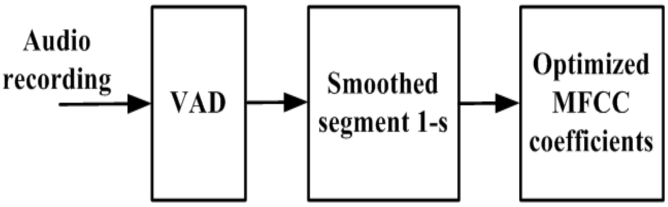


Figure 1: The procedure of signal processing to obtain optimized MFCCs coefficients.

2. The optimized MFCCs

Deepfake audio signals often consist of similar features to the original human voice. However, distinguishing ability is a difficult problem to advance in deep learning approaches in detecting and classifying deepfake. Hence, distinct features can significantly affect the model's predictive ability and effectiveness. In the frequency domain, it is observed that audio signals can provide us helpful characteristics in detection and classification of deepfake audio. For this purpose, the author uses optimized Mel-frequency Cepstral Coefficient (MFCCs), which play an important role in speech recognition, because MFCCs extraction is more computational demanding than Zero-Crossing Rate or energy calculation. MFCCs remain feasible for embedded acoustic devices with integrating various signal processing algorithms.

For audio data, for each segment of audio, the short-term energy is calculated as:

$$E = \frac{1}{N} \sum_{n=0}^{N-1} x[n]^2 \tag{1}$$

where $x[n]$ denote the normalized audio sample and N is the length of considered frame. The computed energy is the converted in decibel as the formulation:

$$E_{dB} = 10 \log_{10}(E + \epsilon) \tag{2}$$

where ϵ is small additive constant to prevent the logarithmic singularities.

The MFCC feature extraction can be computed as:

1) Applying a Hamming window and transforming it using a 1024-point FFT to achieve magnitude and phase value. The number of FFT points can be used for enhancing the accuracy of detection. In this article, the author used 1024-point to achieve a better balance between signal processing time and performance.

2) A precomputed mel filterbank for 44,1 kHz sampling and 1024-point is implemented for obtaining the approximation of human auditory perception.

3) Computation logarithm and Discrete Cosine Transform (DCT) to take into account the first 64 MFCC coefficients.

The author's idea is using an efficient Voice Activity Detection (VAD) for determining the frame with presence/absence of speech component for calculating exactly MFCCs. On the assumption that the speech component often lies on frequency range $0 - 3400$ Hz, and noise distributes on higher bandwidth frequencies. The author proposed computing the ratio spectral energy between bandwidth $0 - 3400$ and $3400 - 22100$ Hz.

$$VAD = \frac{E_{0-3400(Hz)}}{E_{3400-22100(Hz)}} \tag{3}$$

If the VAD greater than a determined threshold γ, the frame contains the speech component and otherwise.

$$\begin{cases} VAD > \gamma & \text{speech} \\ VAD \leq \gamma & \text{noise} \end{cases} \tag{4}$$

Table 1
The CNN, LSTM, GAN and MLP Network Architecture

Models	Configuration
CNN	2 x Conv(64) - ReLU - AP - Dropout(0.3)
LSTM	2 x BiLSTM(64)-ReLU - Dropout(0.3)
GAN	2 x Conv(64) - ReLu - Dropout(0.3)
MLP	1 x Dense(64) - ReLu

The scheme of determining speech/noise frame is illustrated in Figure 1. With the goal of estimating a reasonable threshold to avoid false VAD where there is only ambient noise. The author's procedure can be expressed as:

1) Select the first 5-10 frames at start-up with the assumption that no speech component is available during this period.

2) Compute acoustic features, such as spectral, energy and MFCCs.

3) Calculate the properties of surrounding noise by taking the average of the above parameters.

4) Set adaptive thresholds based on the background noise. Save the results in a global variable VAD to use during the entire signal processing.

During the speech frame, the pre-emphasis of MFCC is performed according to the equation:

$$y[n] = x[n] - \alpha x[n-1] \tag{5}$$

where $y[n]$ is the output, $x[n]$ is the input and $x[n-1]$ is the previous value of $x[n]$, and pre-emphasis parameter is $\alpha = 0.9$.

In this section, the MFCCs were precisely computed during the entire recording file and it can improve the accuracy of detection deepfake.

3. The author's proposed method and other classification models

Convolutional Neural Networks (CNNs) have been commonly implemented for detection tasks due to its capability to process spatial dependencies in audio spectrograms. The scholars have designed CNN architectures for extracting the discriminative features from observed audio signals.

Long Short-Term Memory (LSTM) , which is an efficient technique for resolving temporal dependencies in sequential data, is suitable for analyzing and processing audio signals. LSTM adepts operate at learning long-range dependencies and can be applied to various tasks such as speech recognition, source separation, VAD and deepfake audio detection.

To address the audio deepfake detection, Generative Adversarial Networks (GAN)s have been used for training detection models through generating synthetic audio samples.

Multi-Layer Perceptron (MLP) is an efficient solution for classification problems. A multilayer perceptron, through layers, can effectively outperform the relevant features from data and tune the parameter of the models for optimal predictions. In the MLP model, there are at least three levels: an input layer, a hidden layer of calculation nodes and an output layer of processing nodes.

The author's idea is using the optimized MFCCs feature to increase the effectiveness of audio deepfake detection by using the above ML algorithm. The final decision is made based on the majority of obtained results. The author uses Librosa library [9] to convert real/fake audio files and each audio recording is divided into small 1 second slices. In this stage, MFCCs coefficients are computed by applying fast Fourier transform, logarithmic and discrete cosine transform.

The proposed system (ProSys + MFCC) was presented in Figure 2.

As an individual model operates on 1 - second audio segment, the overall predicted probability of a certain audio recording segment is determined by averaging of estimated probabilities over entire audio file. If we denote $\boldsymbol{g}^{(L)} = [g_1^{(L)} \quad g_2^{(L)} \quad \ldots \quad g_M^{(L)}]$ with M means the category number of the L-th out of L 1-second segments in one audio file. The computed probability of considered audio recording is

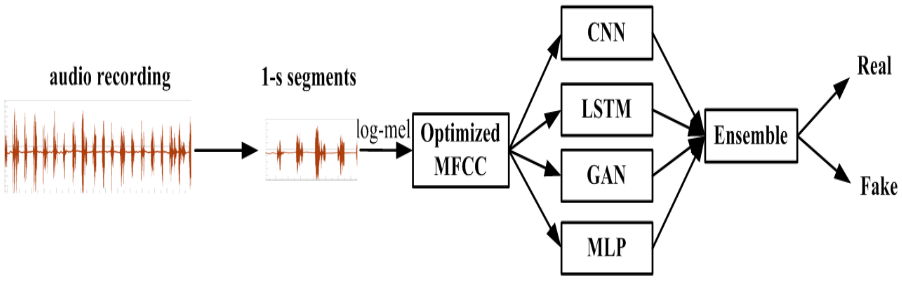
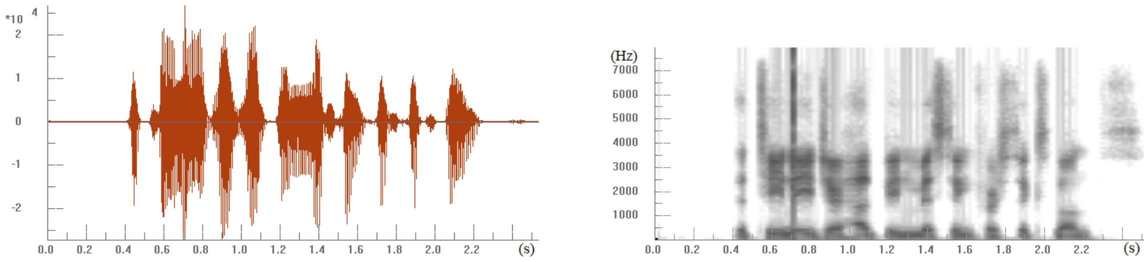


Figure 2: The procedure of signal processing to obtain optimized MFCCs coefficients.



(a)

(b)

Figure 3: The waveform (a) and spectrogram (b) of entire recording in Logic Access of AVSpoofing 2021.

calculated by averaging the classification probability which denoted as $\overline{\boldsymbol{g}}^{(L)} = [\overline{g}_1^{(L)} \quad \overline{g}_2^{(L)} \quad \dots \quad \overline{g}_M^{(L)}]$ where:

$$\overline{g}_M = \frac{1}{L}\sum_{n=1}^{L} g_i^{(n)} \quad 1 \leq i < M \tag{6}$$

For ensembel of results from different models, the author demonstrated experiment on the individual models, then achieved the predicted probability as $\hat{\boldsymbol{g}}_s = [\overline{g}_{s_1} \quad \overline{g}_{s_2} \quad \dots \quad \overline{g}_{s_M}]$ of S individual evaluated model. Next, the predicted probability after MEAN fusion $\hat{\boldsymbol{g}}_{mean} = [\hat{g}_1 \quad \hat{g}_2 \quad \dots \quad \hat{g}_M]$ is derived by:

$$\hat{g}_c = \frac{1}{S}\sum_{s=1}^{S} \tag{7}$$

After all, the predicted label $\hat{y}$ can be expressed as:

$$\hat{y} = argmax(\hat{g}_1, \hat{g}_2, ..., \hat{g}_M) \tag{8}$$

4. Experiments

In this experiment, the author evaluated the suggested models on the Logic Access dataset of ASVspoofing 2019 challenge with real and fake audio samples, which were generated by an AI-based generative system. Logic Access can be categorized into three subsets 'Train' (22800/2580), 'Develop' (22296/2548) and 'Evaluation' (63882/7355) (fake sample/real sample). The author utilized the 'Train' subset for the training model, then performed the 'Develop' subset. Finally, the models are tested on the 'Evaluation' subset. The promising results were reported through Equal Error Rate (ERR), Accuracy, F1 score and AuC score. The author's approach using the optimized MFCC has the advantage of improving the capability of audio fake detection.

The author's method is comparing evaluation of ProSys + optMFCC with each ML algorithm with normal calculation MFCC.

Table 2
Performance Comparison Among ML Algorithm on Logic Access in AVSpoofing 2021

Models	Acc	F1	AuC	ERR
CNN	0.82	0.84	0.86	0.15
RNN	0.84	0.85	0.83	0.14
GAN	0.84	0.86	0.89	0.17
MLP	0.77	0.81	0.87	0.22
ProSys + OptMFCC	**0.86**	**0.86**	**0.94**	**0.88**

The model ProSys + optMFCC with optimized MFCCs coefficients and ensembles of all ML algorithms has shown the increased performance in Accuracy, F1 score, AuC and ERR. With an appropriate VAD, the author's method can determine whether a frame with presence/absence speech component to exactly compute the necessary MFCC coefficients. With individual ML algorithms, the input frame is not smoothed, therefore the MFCC feature does not bring all characteristics of human voice, which plays an important role in the decision of detecting audio fake samples. Consequently, ProSys + optMFCC gave us better accuracy, F1 score, AuC. Beside, the ensemble method has improved the final result of ProSys + optMFCC. The promising achieved ERR 0.08 in comparison with CNN(0.15), RNN (0.14) and GAN (0.17). These findings confirmed that the diverse features via ensemble multiple ML algorithms, which are based on spectrograms, substantially improves the overall evaluation compared to single technique.

5. Conclusion

This paper has described an efficient approach of optimum calculation of MFCCs coefficients and ensemble techniques, which is based on spectrogram features. The appealing properties of the author's approach is utilizing the VAD to exactly choose a frame with presence of speech component, smoothed and calculated 64 first coefficients. Beside, an effective ensemble technique, which based on spectral features, outperformed better accuracy and ERR on AVSspoofing 2021 database. The numerical results have confirmed the ability of the suggested system in addressing many complex problems.

Declaration on Generative AI
The author(s) have not employed any Generative AI tools.

References

[1] A. Abbasi, A. R. R. Javed, A. Yasin, Z. Jalil, N. Kryvinska and U. Tariq, "A Large-Scale Benchmark Dataset for Anomaly Detection and Rare Event Classification for Audio Forensics," in IEEE Access, vol. 10, pp. 38885-38894, 2022, doi: 10.1109/ACCESS.2022.3166602.

[2] A. R. Javed, W. Ahmed, M. Alazab, Z. Jalil, K. Kifayat and T. R. Gadekallu, "A Comprehensive Survey on Computer Forensics: State-of-the-Art, Tools, Techniques, Challenges, and Future Directions," in IEEE Access, vol. 10, pp. 11065-11089, 2022, doi: 10.1109/ACCESS.2022.3142508.

[3] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention," in Proc. Interspeech, Hyderabad, 2018, pp. 681–685. DOI:10.21437/Interspeech.2018-2279.

[4] S. Pradhan, W. Sun, G. Baig, and L. Qiu, "Combating replay attacks against voice assistants," Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol., vol. 3, no. 3, pp. 1–26, Sep. 2019. DOI:10.1145/3351258.

[5] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems,"

2011 Carnahan Conference on Security Technology, Barcelona, Spain, 2011, pp. 1-8, doi: 10.1109/CCST.2011.6095943.

[6] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, “The ASVSPOOF 2017 challenge: Assessing the limits of replay spoofing attack detection,” in Proc. 18th Annu. Conf. Int. Speech Commun. Assoc., 2017, pp. 2–6.

[7] J. Frank and L. Schönherr, “WaveFake: A data set to facilitate audio deepfake detection,” 2021, arXiv:2111.02813.

[8] M. Hassaballah, M. A. Hameed, and M. H. Alkinani, “Introduction to digital image steganography,” in In book: Digital Media Steganography Principles, Algorithms, and Advances (pp.1-15), DOI:10.1016/B978-0-12-819438-6.00009-8