electronics

Article

Deepfake Voice Detection: An Approach Using End-to-End Transformer with Acoustic Feature Fusion by Cross-Attention

Liang Yu Gong *,† and Xue Jun Li *,†

Department of Electrical and Electronic Engineering, Auckland University of Technology, Auckland 1010, New Zealand

* Correspondence: liangyu.gong@autuni.ac.nz (L.Y.G.); xuejun.li@aut.ac.nz (X.J.L.)

† These authors contributed equally to this work.

Abstract: Deepfake technology uses artificial intelligence to create highly realistic but fake audio, video, or images, often making it difficult to distinguish from real content. Due to its potential use for misinformation, fraud, and identity theft, deepfake technology has gained a bad reputation in the digital world. Recently, many works have reported on the detection of deepfake videos/images. However, few studies have concentrated on developing robust deepfake voice detection systems. Among most existing studies in this field, a deepfake voice detection system commonly requires a large amount of training data and a robust backbone to detect real and logistic attack audio. For acoustic feature extractions, Mel-frequency Filter Bank (MFB)-based approaches are more suitable for extracting speech signals than applying the raw spectrum as input. Recurrent Neural Networks (RNNs) have been successfully applied to Natural Language Processing (NLP), but these backbones suffer from gradient vanishing or explosion while processing long-term sequences. In addition, the cross-dataset evaluation of most deepfake voice recognition systems has weak performance, leading to a system robustness issue. To address these issues, we propose an acoustic feature-fusion method to combine Mel-spectrum and pitch representation based on cross-attention mechanisms. Then, we combine a Transformer encoder with a convolutional neural network block to extract global and local features as a front end. Finally, we connect the back end with one linear layer for classification. We summarized several deepfake voice detectors' performances on the silence-segment processed ASVspoof 2019 dataset. Our proposed method can achieve an Equal Error Rate (EER) of 26.41%, while most of the existing methods result in EER higher than 30%. We also tested our proposed method on the ASVspoof 2021 dataset, and found that it can achieve an EER as low as 28.52%, while the EER values for existing methods are all higher than 28.9%.

Keywords: end-to-end; transformer; cross attention; feature fusion; supervised learning; deepfake voice recognition

Received: 8 April 2025
Revised: 7 May 2025
Accepted: 13 May 2025
Published: 16 May 2025

Citation: Gong, L.Y.; Li, X.J. Deepfake Voice Detection: An Approach Using End-to-End Transformer with Acoustic Feature Fusion by Cross-Attention . *Electronics* **2025**, *14*, 2040. https://doi.org/10.3390/electronics14102040

1. Introduction

Deepfakes pose significant risks in today's digital world [1]. By creating highly realistic but fake audio, video, or images, they can spread misinformation, damage reputations, and manipulate public opinion. Deepfakes have been used for political propaganda, financial fraud, and cyberbullying. In some cases, they have even been exploited to produce non-consensual explicit content, leading to severe emotional and psychological harm to victims. The challenge of telling real from fake undermines trust in digital media and puts the integrity of information at risk.

Detecting deepfakes is crucial to maintaining trust in digital communication and media. Practical detection tools help to identify manipulated content early on, preventing its spread and limiting the damage it can cause. This is especially vital for protecting individuals, institutions, and democratic processes. As deepfake technology becomes more sophisticated, developing advanced detection methods using AI, blockchain verification, or digital watermarking is essential for avoiding potential misuse and ensuring online safety.

Deepfake voices pose serious threats by enabling realistic impersonation, which can be used for scams, fraud, and spreading misinformation, undermining trust in voice-based communication. Therefore, detecting deepfake voices is as crucial as detecting deepfake videos/images [2,3]. Currently, forgery speeches can be summarized in two main categories, as shown in Figure 1: text-to-speech (TTS) and voice conversion (VC). Unlike traditional voice splicing and editing, digital voice manipulations can generate much smoother waveforms with the development of acoustic feature extraction and vocoders. Therefore, deepfake voice technology has various applications in society; for example, TTS can be utilized in the movie industry, helping people with voice impairments and protecting speakers' privacy and identities by changing their timbre and pitch [4]. On the other hand, VC systems aim to convert source speeches to specific target speeches without changing their linguistic contents [5]. Traditional VC systems use high-quality vocoders, such as WORLD [6], to extract acoustic features separately, including different frequency spectrums and fundamental frequencies (F0). A waveform generator is usually applied to generate the corresponding waveform from the extracted features.

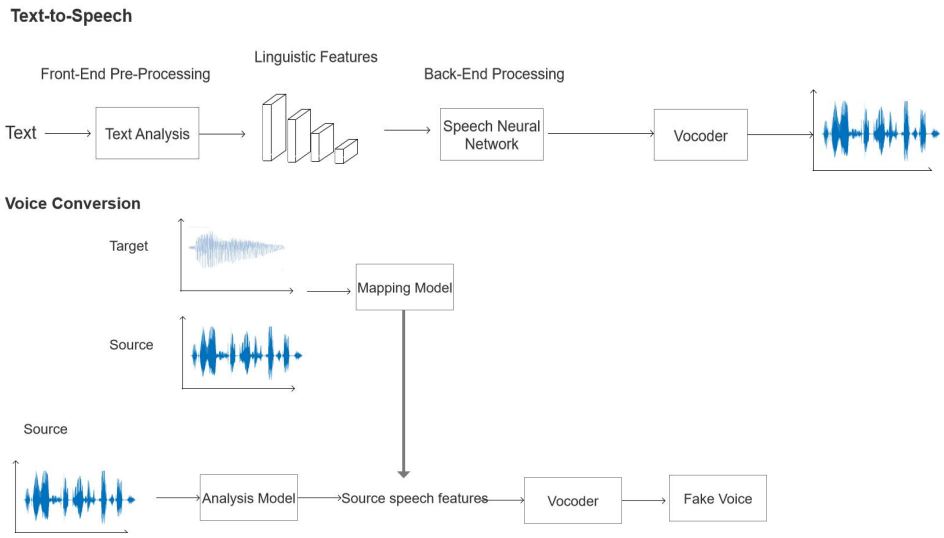

Figure 1. A common category of digital voice manipulation.

Based on those excellent voice forgery models and open-source audio forgery tools, high-quality logistic attack audio can be generated and has many illegal applications that make the audio hard to distinguish by Human Auditory Systems (HASs). There are numerous cases illustrating that deepfake voice technology poses a threat to our daily lives. For example, deepfake voice technology can deceive speaker verification systems, and some fraudsters have exploited this vulnerability to conduct telephone fraud. Thus, there is now an urgent need to develop a reliable deepfake voice detection system with strong generalization abilities to classify the authenticity of the audio.

Currently, several difficulties remain under-explored in deepfake voice detection. Firstly, deepfake voice detectors always require large amounts of training data to extract the different acoustic features of real and fake audio. This leads to significant challenges in the training process. Secondly, existing deepfake voice recognition systems have weak gen-

eralization abilities and cannot perform well in cross-dataset evaluation. Finally, regarding feature extraction, RawNet [7] directly utilizes raw audio instead of other acoustic features as the input to start classification work. The more varied the input acoustic features, the better the performance of models on the cross-dataset evaluation.

The main objective of this study is to design a more generalized deepfake voice detector to classify the real and forged audio. In extracting acoustic features, we analyzed them in terms of the complementarity of speech characteristics, as the Mel-spectrogram and F0 signal effectively represent phonemes and pitch, respectively. In addition, the Transformers with CNN modules are rarely applied in deepfake voice detection, so we proposed a fusion of the acoustic features using the cross attention mechanism and design a deepfake voice detector in combination with the Transformer and CNN block. Therefore, we propose an end-to-end (E2E) Transformer-based deepfake voice detector with the fusion of Mel-spectrogram and F0. The complete architecture of the proposed Deepfake voice detector is illustrated in Figure 2. More specifically, the acoustic feature fusion employs a self-attention mechanism to process the Mel-spectrogram, a linear projection layer to handle the F0 signals as pitch representations, and ultimately integrates these two acoustic features adaptively using cross-attention.To reduce training time, we employed a transfer learning approach by loading partial weights from the pre-trained Conformer model [8] weights and fine-tuning it on the ASVspoof2019 [9] dataset. The main contributions of this work are threefold:

- To address the challenge of enriching the model's input through the fusion of diverse acoustic features, we propose a PreNet architecture that incorporates an attention-based feature fusion method to adaptively combine the pitch representation and Mel-spectrogram features.
- To enhance acoustic feature extraction, the original Transformer integrates convolutional neural network (CNN) modules to separately capture local and global features. Unlike the Conformer architecture, we position the convolutional blocks at the beginning of the extractor workflow, followed by a feed-forward network (FFN) to adjust the input tensor's dimensionality.
- We integrate the PreNet with the custom Transformer-based extractor as the front-end and connect it to a linear layer, forming an end-to-end Deepfake voice detector architecture. Subsequently, we adapted a portion of the pre-trained Conformer model's weights and fine-tuned the system using the ASVspoof2019 dataset.

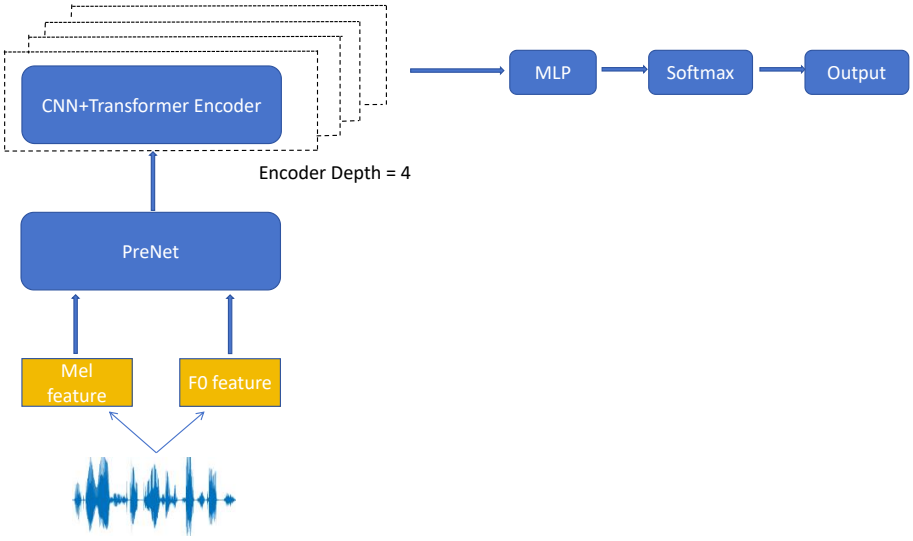

Figure 2. The architecture of the entire proposed deepfake voice detector.

The rest of this paper is organized as follows. Section 2 provides an overview of related work, covering commonly used acoustic features and the Transformer backbone. Section 3 details the proposed method, including the PreNet for fusing Mel-frequency and F0 signals, as well as the complete architecture of a Transformer-based extractor integrated with 1D depthwise CNN modules. Section 4 describes the experimental setup and performance evaluation in comparison with current state-of-the-art (SOTA) models. Finally, Section 5 concludes the paper and discusses directions for future work.

2. Related Work

This section introduces the relative acoustic features, popular backbone architecture, and several deepfake voice detection algorithms.

2.1. Acoustic Features

In the literature, two primary acoustic features are commonly considered for detecting deepfake voices: the Mel-spectrogram and the fundamental frequency (F0). The Mel-spectrogram is a standard audio processing technique that transforms time-domain speech signals into frequency-based representations. It generates non-linear frequency signals based on the Mel scale, which more effectively captures low-frequency band information, making it better suited for HAS, as the human ear is less sensitive to high-frequency speech signals than to low-frequency ones. In speech recognition tasks, signals below 1 kHz are generally considered low-frequency, while those above 3 kHz are regarded as high-frequency. These signals are often related to the pronunciation and clarity of consonants. Typically, a certain amount of high-frequency information is present in human speech. For instance, phonemes such as "/s/" and "/shi/" rely more on high-frequency components [10]. The Mel-spectrogram, which represents the full range of frequencies using the Mel scale, is widely used in acoustic feature extraction. Additionally, the audio signal processed by the Mel-spectrogram becomes two-dimensional (time axis versus Mel-frequency axis), making it well suited to models like RNNs or Transformers [11], which simplify learning for speech conversion models. The Mel-spectrogram requires voice framing with a 25 ms window to divide the audio into smaller segments, ensuring the stability of each acoustic feature. Once the voice frames are generated, Short-Time Fourier Transforms (STFT) are applied to convert time-domain speech into the frequency domain. The Hamming window, used in the pre-processing step, effectively suppresses spectral leakage due to its low side lobes. Finally, designing a set of Mel filters, shown in Equation (1), could convert the original frequency signals to non-linear Mel-frequency. On the contrary, Mel-frequency cepstral coefficients (MFCC) can be explained by Equation (2):

$$Mel(f) = 2595 \cdot \log_{10}(1 + \frac{f}{700}) \tag{1}$$

$$f = 700 \times (10^{\frac{mel}{2595}} - 1) \tag{2}$$

where f represents the original frame's frequency, and mel represents the Mel spectrum correspondingly.

F0 is one of the critical parameters in the speech signal, which refers to the fundamental frequency of the vocal cord vibration. It determines the pitch of the voice and plays an essential role in intonation, stress, and prosody. More specifically, F0 positively correlates with pitch, so several research works utilised F0 to represent the pitch. Since information such as emotions and attitudes can be expressed through the fluctuations of F0, and the fundamental frequencies of voice vary among different individuals, the F0 signal can better reflect the speaker's identity in terms of voice characteristics. Many VC systems

still adopt traditional methods for extracting F0 features, such as YIN [12] or Harvest [6]. These algorithms are computationally efficient and widely used, making them excellent choices for audio pre-processing. However, these two algorithms struggle in noisy speech environments, resulting in a decrease in the accuracy of voice conversion.

2.2. Transformer

RNNs [13] process word embeddings sequentially, one by one, which makes RNN training time-consuming. Additionally, this architecture is constrained by local information and sequential processing, limiting its performance to some extent. If the sequence contains long temporal dependencies, the model may experience gradient vanishing during back-propagation, causing it to forget long-distance information. However, Transformer [11], as a Sequence-to-Sequence (seq2seq) architecture, was published in 2017 for solving long-term dependence problems. It proposed an attention mechanism on word embeddings to focus on global word relationships. In voice conversion models, the Transformer, as shown in Figure 3, typically consists of multiple Encoders and Decoders. Each Encoder contains two sub-layers: the first layer is mainly a multi-head attention module based on Scale Dot production, which enables it to focus on different embedding features, achieving parallel computing. The second sub-layer is an FFN for performing nonlinear computations and dimension adjustment. In addition, each trainable parameter layer is often connected with a shortcut to prevent the degradation of the model's performance. The Decoder components are similar to the Encoder components, but they include an additional masked multi-head attention module to process the target embeddings. The Encoder outputs (K and V) and the Q embeddings from the masked multi-head attention module are then passed to the second sub-layer. Currently, the Transformer architecture is widely used to address audio and NLP problems in various studies. For example, Tacotron 2 [14] integrates the Transformer with the TTS (text-to-speech) task. Additionally, BERT [15], a pre-trained NLP model based on the Transformer encoder, is employed to learn diverse language representations.

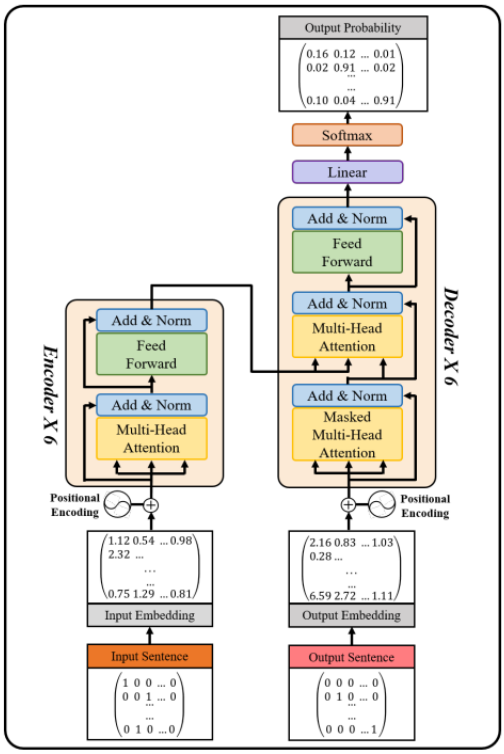

Figure 3. A traditional Transformer architecture; the image is from [16].

2.3. Deepfake Voice Detection

Currently, deepfake voice detection algorithms can be categorized into three main types: traditional acoustic feature-based methods, computer vision-based methods, and end-to-end (E2E) architecture-based detectors. Most deepfake voice detection systems select the Mel-spectrogram as input in the first category. They transfer voice waveforms to feature representations for model learning. For example, Wang et al. [17] proposed a 135-layer dense convolutional network for processing voice transformation (spoofing). It could outperform new benchmark results in cross-dataset evaluation compared with other SOTA models in 2019. Deep4SNet [18] was inspired by computer vision image recognition methods. It converted the voice signal into the corresponding image representation (histograms) and applied image augmentation as pre-processing. They then proposed a CNN-based model and used Dropout to prevent overfitting. E2E approaches have become popular in AI-synthesized voice detection, as these methods use raw voice input instead of relying on additional acoustic feature extraction blocks. For example, RawNet2 [19] bypassed traditional feature extraction methods like Linear-Frequency Cepstral Coefficients (LFCC) extraction. Instead, it takes raw voice signals as input and uses a sinc convolution layer to extract voice features.

3. Proposed Method

Section 3 presents the PreNet designed to extract and fuse the Mel-spectrogram with F0 representation. The complete Transformer-based extractor architecture is presented to explain the process of acoustic feature processing. Finally, one common back-end model with cross-entropy loss is applied for deepfake voice classification.

3.1. PreNet

Our proposed method is an end-to-end (E2E) architecture that uses raw voice signals as inputs. In contrast to other traditional AI-synthesized voice detectors, which typically apply a pre-processing module to convert raw voice into Mel-frequency features for model training, we introduce a PreNet. This PreNet is integrated at the beginning of the model's front-end to process the input waveform and extract the corresponding acoustic features. The workflow is shown in Figure 4, where the primary goal of PreNet is to extract Mel-spectrogram and F0 signals as two acoustic features, encode them, project them to the same 2-dimensional shape, and then apply a cross-attention mechanism to adaptively fuse them into input embeddings.

Since the Mel-spectrogram performs a non-linear mapping of voice frequencies across the entire range and effectively represents phonemes and speech content, it is chosen as the first acoustic feature to extract. Additionally, F0, the fundamental frequency, better reflects the speaker's emotional tone and pitch characteristics, and is positively correlated with pitch level. We extract the F0 feature and fuse it with the Mel-spectrogram to create more comprehensive acoustic features. Specifically, F0 signals are extracted using the YIN algorithm [12] instead of designing an RNN-based F0 extraction model to reduce computational complexity.

Once the original Mel-spectrogram and F0 signals are obtained, we treat the Mel-spectrogram as a 2-dimensional signal with the shape of $[num_{mel}, timesteps]$. We first apply a linear layer to generate the corresponding Query (Q_{mel}), Key (K_{mel}), and Value (V_{mel}), use sine-cosine positional encoding, and then apply an 8-head self-attention module to embed the Mel-spectrogram, as shown in Equation (3):

$$A_{mel} = Softmax(\frac{Q_{mel}K_{mel}^{T}}{\sqrt{d_k}})V_{mel} \quad (3)$$

where $Softmax$ is an activation function; Q_{mel}, K_{mel}, and V_{mel} are the Query, Key, and Value of Mel-spectrogram; and d_k represents the Query dimension which equals d_{model}/num_{heads}.

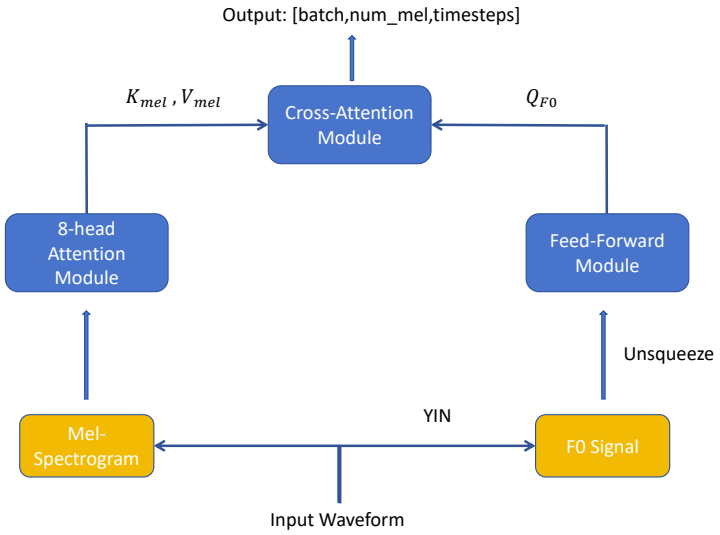

Figure 4. Our proposed cross-attention PreNet, which processes the F0 query representation and Mel key and value representations.

However, F0 signals are 1D signals with the shape of $[timesteps]$, so we add an extra dimension to the last dimension of the original F0 signals, converting them to the shape of $[timesteps, 1]$. Moreover, silent periods during pauses in human speech often result in NaN values in some parts of the F0 signal. Hai et al. [1] opted to remove the silent segments at the input stage, which led to a loss of voice information. To avoid this, we use the traditional linear interpolation method to fill in the NaN values, ensuring the continuity of the F0 signals, rather than replacing the NaN values with zeros. Then, the re-processed voice signals are segmented into voice clips with identical segment length by padding, as shown in Figure 5. After processing the original F0 signals, we adopted a feed-forward network to encode them as pitch representation. The linear progress is shown as Equation (4). Finally, after obtaining the Mel representations and pitch representations, similar to the process of Mel feature extraction, we apply a cross-attention mechanism to adaptively assign balanced weights and fuse the pitch representations with the Mel representations. The combined acoustic embedding is with the same dimension of extracted Mel-spectrogram representations and shown as Equation (5).

$$A_{F0} = \sigma(W_{F0}X_{F0} + b_{F0}) \tag{4}$$

Here, σ represents the sigmoid function, W_{F0} and b_{F0} are weights and bias of linear layer, and X_{F0} represents the F0 2-dimensional tensor.

$$A_{(mel,pitch)} = Softmax(\frac{Q_{F0}K_{mel}^T}{\sqrt{d_k}})V_{mel} \tag{5}$$

Here, Q_{F0} is the Query of pitch representation and K_{mel} and V_{mel} are the Key and Value of Mel-spectrogram.

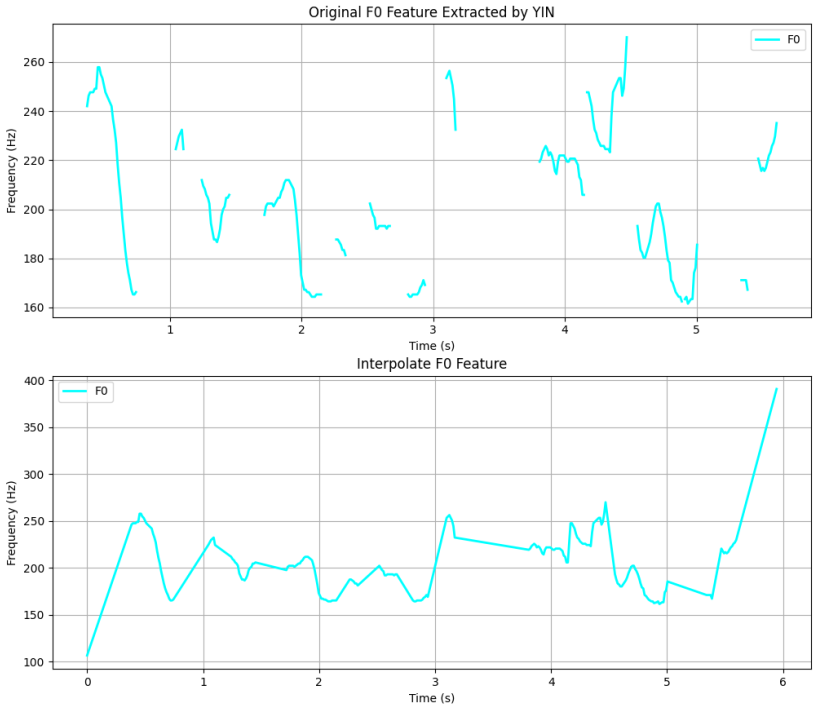

Figure 5. Original F0 signal versus processed F0 signal by padding silence segments in ASVspoof2019.

3.2. Transformer-Based Extractor

Transformers [20,21] have already proven to be a robust backbone in speaker recognition systems and fake audio detection. Thanks to their multi-head attention mechanism, Transformers can extract global features, address long-term dependency issues, and enable parallel processing. However, we strongly believe that local feature extraction is just as crucial as global feature extraction, and CNN modules are highly effective at extracting local features. Based on the structure and design concepts of Conformer [8], we made adjustments to its architecture to combine both local and global features in the acoustic feature extractor. Specifically, the representations obtained from PreNet pass through $N = 3$ 1D depthwise convolutional blocks to extract local features, followed by an FFN, a multi-head self-attention module, another FFN, and layer normalization. Additionally, each trainable layer includes a shortcut connection to prevent model degradation during training. The architecture of the one-layer Transformer-based extractor is shown in Figure 6. The entire encoder of this model is designed with a depth of 4 to ensure that the acoustic feature extractor can effectively capture complex acoustic information, learn long-range sequential dependencies, and gradually extract meaningful features.

To capture local features, the convolutional module is designed as follows. First, a layer normalization step normalizes the input features to enhance training stability and accelerate convergence. Next, a pointwise convolutional layer with a 1×1 kernel adjusts the feature dimensions to meet the computational requirements of 1D depthwise convolution. The Gate Linear Unit (GLU), a non-linear activation function, is applied to improve the model's generalization ability. The core component of the convolutional block is depthwise separable convolution, which was originally introduced in the Xception Network [22]. This method helps save computational costs by reducing the number of parameters. Depthwise separable convolution consists of depthwise convolution and pointwise convolution. Depthwise convolution applies convolution operations to each input channel independently to extract local features, while pointwise convolution uses 1×1 convolution kernels to linearly combine the outputs of the depthwise convolution, thereby fusing information across channels. To further enhance training speed and model generalization, batch normal-

ization is applied after the convolutional layer. Additionally, the Swish activation function is used before the pointwise convolution, as it is smooth and introduces non-linearity. Finally, pointwise convolution is used to replace the MLP layer, ensuring consistency with the input tensor's feature dimensions for the shortcut connection.

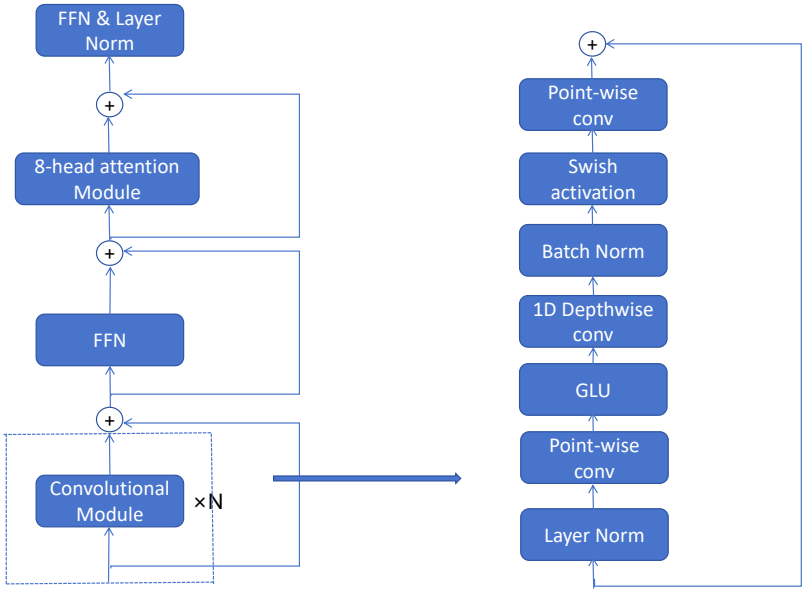

Figure 6. A proposed Transformer-based acoustic feature extractor combined with CNN modules.

3.3. Back-End Model

In the downstream model, we use a standard MLP structure for classification. First, a batch normalization layer is applied to reduce internal covariate shift, stabilize the gradients, and accelerate model convergence. Next, a linear layer with an additional Dropout module is used. The linear layer projects the input dimension to the hidden layer dimension, enhancing the model's representation ability, while the Dropout module helps prevent overfitting. Finally, another linear layer adjusts the output dimension to match the number of predicted classes, and a Softmax function is applied for non-linear performance. The details of the MLP components are shown in Figure 7.

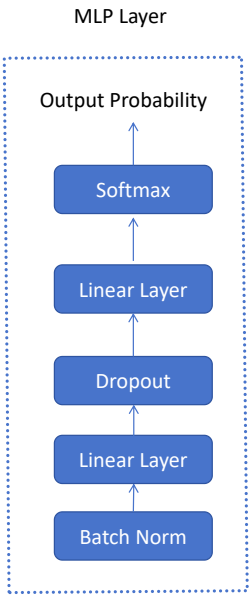

Figure 7. An MLP layer for classification as downstream model.

4. Experimental Evaluation

In this section, we present the experimental datasets and setup parameters. Additionally, we compare several benchmark models with our proposed method in both an in-dataset experiment and a cross-dataset experiment.

4.1. Dataset

We conducted our training experiment on the Logistic Access ("LA") subset of ASVspoof 2019 [9]. This dataset, an extension of previous ASV spoof challenges, is designed to detect spoofed and genuine audio, with spoof data generated using common text-to-speech and voice conversion techniques. The LA dataset consists of three subsets: the training set, which includes 20 speakers (8 males and 12 females); the development set, with 10 speakers (4 males and 6 females); and the testing set, which contains 48 speakers (21 males and 27 females). Specifically, the LA training dataset includes 6 spoofing algorithms, with 2580 bonafide samples and 22,800 spoof samples. The development subset has 2548 bonafide samples and 22,296 spoof samples. Finally, the evaluation set of ASVspoof 2019 contains 7335 bonafide and 63,882 spoof samples. For cross-dataset testing, we use the ASVspoof 2021 [23] evaluation subset to assess our model's generalization capability. This dataset includes 67 speakers, 13 spoofing algorithms, 163,114 fake audios, and 18,452 real audios.

4.2. Experimental Setups

This section outlines the technical details of the proposed model. We use the "Adam" optimizer throughout the entire training process, with a learning rate of 0.0001. The number of convolutional modules for extracting local features is set to 3, the depth of the Transformer-based extractor is set to 4, and the number of training epochs is set to 60. The experimental setup details for the acoustic feature fusion PreNet are provided in Table 1.

Table 1. The experimental setup details of PreNet.

Experimental Parameters	Details
Sample rate (SR)	16,000
Hop length	256
Mel-frequency channels	80
Multi-head attention number	8
Dropout rate	0.2
Model dimension	1024
Feed-forward network dimension	2048

4.3. Evaluation Results

In this section, we performed in-dataset experiments on the ASVspoof 2019 dataset to evaluate our model's performance on the processed silence segments. We then applied the ASVspoof 2021 dataset to assess our model's generalization ability. Specifically, in the in-dataset evaluation experiments, we selected five representative fake voice detectors (models published in 2021 and 2022) to compare their in-dataset Equal Error Rate (EER) performances. Since we interpolated the silence segments of raw audio, the benchmark models were also evaluated with the removal or processing of the silence segments. The selected baseline deepfake voice detectors were RawNet2 [19], AASIST [24], Fastaudio [25], RawGAT-ST [26], and MTLISSD [27]. Based on the results of SiFsafer's experiments [1], it was concluded that the deepfake voice detection systems are more vulnerable in processing silence segments. The validation results of the model after processing silent segments more effectively demonstrate the model's robustness in in-dataset validation. Additionally, deep-

fake voice detection in silence segment processing will present more challenges in future work. As a result, this comparison will be more compelling in future detection efforts.

We tested our proposed method on the ASVspoof 2019 LA development and evaluation sets as an in-dataset experiment by calculating the EER results shown in Table 2. Among the compared benchmark models, our model achieved the lowest EER value in both subsets. However, all other models, except for RawNet2, resulted in EER values higher than 30%. The compared models are all supervised learning methods using labeled data, as our model also heavily relies on a large amount of labeled data. This demonstrates that our proposed network can effectively extract deepfake voice acoustic features to some extent. Next, we used the ASVspoof 2021 evaluation dataset to assess our model's generalization ability. The cross-dataset results are shown in Table 3. Our result, with an EER of 28.52%, is slightly lower than Fastaudio's result. Since the cross-dataset evaluation EERs are all above 28%, it indicates that the generalization ability of supervised learning-based deepfake voice detectors still has significant potential.

Table 2. The EER results of selected models in ASVspoof 2019 LA dev and eval subset. All the silent segments are processed.

Detectors	Dev EER	Eval EER
RawNet2	10.28%	28.96%
AASIST	14.40%	30.96%
Fastaudio	12.76%	32.56%
MTLISSD	27.31%	37.56%
RawGAT-ST	17.39%	32.28%
Our model	9.27%	26.41%

Table 3. The EER results of selected models in ASVspoof 2021 LA eval subset.

Detectors	Eval EER
RawNet2	34.39%
AASIST	32.15%
Fastaudio	28.93%
MTLISSD	43.75%
RawGAT-ST	49.64%
Our model	28.52%

5. Conclusions and Future Work

This study presents an end-to-end (E2E) deepfake voice detection framework that incorporates a cross-attention mechanism to integrate multiple acoustic representations. The proposed architecture comprises a Transformer encoder augmented with stacked 1D depthwise separable convolutional blocks to enable effective extraction of both global contextual and local acoustic features. The model is trained in a supervised manner using the cross-entropy loss function. To address the issue of missing pitch data due to silent segments, linear interpolation is employed as a pre-processing strategy, which increases robustness while introducing additional complexity to the detection task. Experimental results demonstrate that the proposed method achieves superior performance compared to existing baselines, attaining a lower Equal Error Rate (EER) in both in-dataset and cross-dataset evaluations. Specifically, the model yields an in-dataset EER of 26.41%, and although its cross-dataset EER is marginally higher than that of Fastaudio, the results underscore the model's effectiveness in generalizing across datasets. Future work will focus on enhancing the generalization capabilities and optimizing the model's performance further.

Future work will primarily explore three key directions. First, we plan to investigate alternative model training paradigms, including the evaluation of different loss functions and composite loss strategies, to further enhance model performance. Given that supervised learning approaches are inherently dependent on large volumes of labelled data, they often suffer from class imbalance issues, where predictions are biased toward the dominant class. While techniques such as data augmentation and resampling can partially mitigate this, transitioning to self-supervised learning paradigms may offer a more principled and scalable solution by leveraging unlabeled data to learn robust representations.

Second, the current study is limited to English-language datasets—namely, ASVspoof 2019 and the deepfake voice subset. To assess the generalizability of the proposed method, future efforts will include extending evaluation to multilingual datasets. This will facilitate a comprehensive analysis of cross-linguistic patterns in synthetic speech generation and detection, and help identify language-invariant acoustic cues associated with deepfake voice forgeries.

Finally, we aim to evaluate model performance on audio data that includes unprocessed silent segments, as such scenarios are more representative of real-world conditions. In particular, we will focus on enhancing the robustness of F0 feature extraction by improving the treatment of NaN values introduced by the YIN algorithm during silent intervals. Linear interpolation will be re-evaluated in favor of more advanced imputation or modeling techniques, and empirical experiments will be conducted to quantify the impact of silent segment handling on detection accuracy.

Author Contributions: Conceptualization, L.Y.G. and X.J.L.; methodology, L.Y.G. and X.J.L.; software, L.Y.G.; validation, L.Y.G.; formal analysis, L.Y.G.; investigation, L.Y.G.; resources, L.Y.G. and X.J.L.; data curation, L.Y.G. and X.J.L.; writing—original draft preparation, L.Y.G. and X.J.L.; writing—review and editing, L.Y.G. and X.J.L.; visualization, L.Y.G. and X.J.L.; supervision, X.J.L.; project administration, X.J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hai, X.; Liu, X.; Tan, Y.; Liu, G.; Li, S.; Niu, W.; Zhou, R.; Zhou, X. What's the Real: A Novel Design Philosophy for Robust AI-Synthesized Voice Detection. In Proceedings of the 32nd ACM International Conference on Multimedia, MM '24, Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 6900–6909.
2. Gong, L.Y.; Li, X.J.; Chong, P.H.J. Swin-Fake: A Consistency Learning Transformer-Based Deepfake Video Detector. *Electronics* **2024**, 13, 3045. [CrossRef]
3. Gong, L.Y.; Li, X.J. A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges. *Electronics* **2024**, *13*, 585. [CrossRef]
4. Gao, Y.; Singh, R.; Raj, B. Voice Impersonation Using Generative Adversarial Networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2506–2510.
5. Stylianou, Y.; Cappe, O.; Moulines, E. Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech Audio Process.* **1998**, *6*, 131–142. [CrossRef]
6. Morise, M.; Yokomori, F.; Ozawa, K. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Trans. Inf. Syst.* **2016**, *99-D*, 1877–1884. [CrossRef]
7. Wang, C.; Yi, J.; Tao, J.; Zhang, C.; Zhang, S.; Fu, R.; Chen, X. TO-Rawnet: Improving RawNet with TCN and Orthogonal Regularization for Fake Audio Detection. *arXiv* **2023**, arXiv:2305.13701.
8. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100.

9. Wang, X.; Yamagishi, J.; Todisco, M.; Delgado, H.; Nautsch, A.; Evans, N.; Sahidullah, M.; Vestman, V.; Kinnunen, T.; Lee, K.A.; et al. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Comput. Speech Lang.* **2020**, *64*, 101114. [CrossRef]
10. Vitela, A.D.; Monson, B.B.; Lotto, A.J. Phoneme categorization relying solely on high-frequency energy. *J. Acoust. Soc. Am.* **2015**, *137*, EL65–EL70. [CrossRef] [PubMed]
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
12. Cheveigné, A.; Kawahara, H. YIN, A fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* **2002**, *111*, 1917–1930. [CrossRef] [PubMed]
13. Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent Neural Network Regularization. *arXiv* **2014**, arXiv:1409.2329.
14. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
15. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019. [CrossRef]
16. Kim, J.W.; Jung, H.Y.; Lee, M. Vocoder-free End-to-End Voice Conversion with Transformer Network. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Piscataway, NJ, USA, 19–24 July 2020; pp. 1–8.
17. Wang, Y.; Su, Z. Detection of Voice Transformation Spoofing Based on Dense Convolutional Network. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2587–2591.
18. Ballesteros, D.M.; Rodriguez-Ortega, Y.; Renza, D.; Arce, G. Deep4SNet: Deep learning for fake speech classification. *Expert Syst. Appl.* **2021**, *184*, 115465. [CrossRef]
19. Tak, H.; Patino, J.; Todisco, M.; Nautsch, A.; Evans, N.; Larcher, A. End-to-End anti-spoofing with RawNet2. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6369–6373.
20. Liu, A.; Yang, S.W.; Chi, P.H.; Hsu, P.c.; Lee, H.y. Mockingjay: Unsupervised Speech Repre- sentation Learning with Deep Bidirectional Transformer Encoders. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6419–6423.
21. Zhang, Z.; Yi, X.; Zhao, X. Fake Speech Detection Using Residual Network with Transformer Encoder. In Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security, IHMMSec '21, Virtual, 22–25 June 2021; pp. 13–22.
22. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
23. ASVspoof 2021 Dataset Download Url. Available online: https://www.kaggle.com/datasets/serjkalinovskiy/asvspoof2021-df (accessed on 7 June 2024).
24. Jung, J.w.; Heo, H.S.; Tak, H.; Shim, H.j.; Chung, J.S.; Lee, B.J.; Yu, H.J.; Evans, N. AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 6367–6371.
25. Fu, Q.; Teng, Z.; White, J.; Powell, M.; Schmidt, D.C. FastAudio: A Learnable Audio Front-End for Spoof Speech Detection. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022.
26. Tak, H.; Jung, J.W.; Patino, J.; Kamble, M.; Todisco, M.; Evans, N. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. *arXiv* **2021**, arXiv:2107.12710.
27. Mo, Y.; Wang, S. Multi-Task Learning Improves Synthetic Speech Detection. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 6392–6396.