Crimson Publishers
Wings to the Research

Mini Review

A New Method for End-to-End Spoof Speech Detection

Jianwu Zhang* and Chen Jia

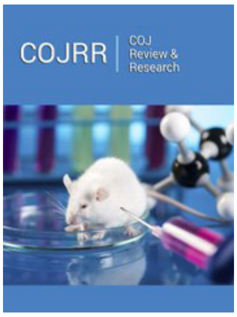
School of Communication Engineering, Hangzhou Dianzi University, China

Abstract

With the rapid development of speech synthesis and speech conversion technology, methods of spoof speech detection still have problems such as low spoof detection accuracy and poor generality. Therefore, an end-to-end spoof detection method based on context information and attention feature was proposed. Based on Deep Residual Shrinkage Network (DRSN), this method used the Dual-branch Context information Coordination fusion Module (DCCM) to aggregate rich context information, and fused features based on Coordinate Time-Frequency Attention (CTFA) to obtain cross-dimensional interaction features with context information, thus maximizing the potential of capturing artifacts.

Keywords: Spoof speech detection; Context information; Attention feature; End-to-End; Artifacts

ISSN: 2639-0590



***Corresponding author:** Jianwu Zhang, School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, China

Submission: 📅 February 13, 2023
Published: 📅 May 5, 2023

Volume 4 - Issue 4

How to cite this article: Jianwu Zhang* and Chen Jia. A New Method for End-to-End Spoof Speech Detection. COJ Rev & Res. 4(4). COJRR. 000594. 2023.
DOI: 10.31031/COJRR.2023.04.000594

Introduction

Automatic Speaker Verification (ASV) systems are used as an identification technology to verify the identity of a speaker from a speech signal [1]. The ASV system verification process does not require any face-to-face contact [2], does not pose discomfort and health risks to the user, but makes the system vulnerable to spoofing attacks.

The framework of currently used anti-spoofing methods consists mainly of front-end feature extraction and back-end classification, where the manual acoustic features generated at the front-end are fed into a back-end classifier. Manual acoustic features can be flawed in detecting invisible attacks, so work has been done to propose an End-to-End (E2E) solution that operates directly on the original audio waveform [3], a solution that effectively avoids the limitations imposed by manual acoustic features. Studies [4,5] have shown that E2E systems outperform classical spoof detection systems, but the results show that there is still much room for improvement.

In the Logical Access (LA) scenario of ASV spoof 2019 [6], synthetic speech spoofing attacks mainly take the form of speech synthesis and speech conversion. Artefacts used to indicate spoofing attacks are called spoofing artefacts, and the artefacts often depend on the nature of the attack and the particular attack algorithm. In the ASV spoof 2021 [7] LA scenario, real speech and spoofed speech are transmitted over various information networks with unknown codecs and transmissions. When voice data is transmitted across information systems, there may be some interfering changes in the transmission channel subjecting spoofing artefacts in the data to unknown codecs and transmissions, increasing the difficulty of spoofing detection and thus the performance requirements for spoofing detection systems. In synthetic speech detection, spoofing artefacts are used to distinguish real speech from spoofed speech and exist mainly in specific temporal and spectral intervals with highly distinguishable temporal and frequency features, but there is currently no better method to capture spoofing cues that can exist between the time and frequency domains. Although the detection performance of existing methods has all improved compared to traditional methods, with the development of various high-quality spoofing attacks, existing spoofing detection methods still lack effectiveness and generality against unknown spoofing attacks.

Based on the original audio waveform, we propose an E2E context information and attention feature fusion network (CAFNet). The structure is shown in Figure 1. The network is based on the Deep Residual Shrinkage Network (DRSN) that uses the Dual-branch Context information Coordination fusion Module (DCCM) to aggregate rich context information and fuse features based on coordinate Time-Frequency Attention (CTFA) to obtain cross-dimensional interaction features with context information to maximize the potential for artefact capture.

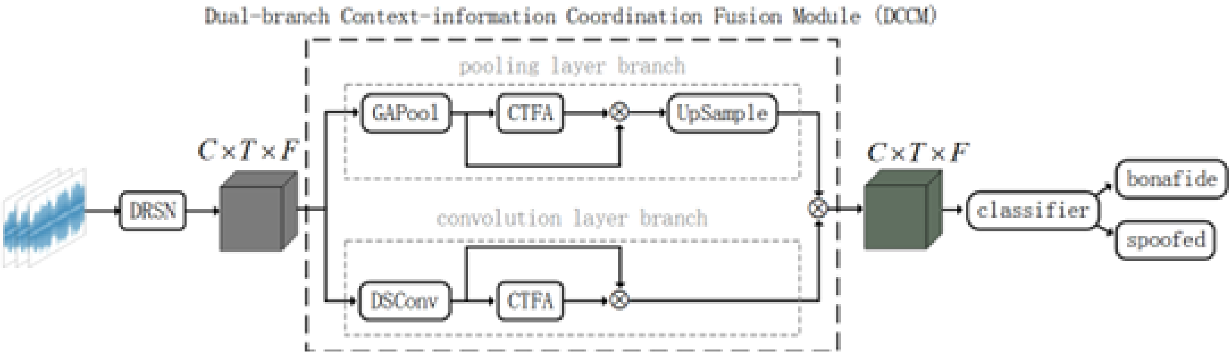


Figure 1: The structure of CAFNet.

DCCM consists of both convolutional layer and pooling layer branches, in which this structure is shown in Figure 2. High-level features are fed into the convolutional and pooling layers respectively, which are convolved and pooled to obtain more contextual information, and then the context-informed features are fused with attention-based features to yield context-informed attention features to acquire rich contextual information and coordinate cross-dimensional interactions of discriminative cues.

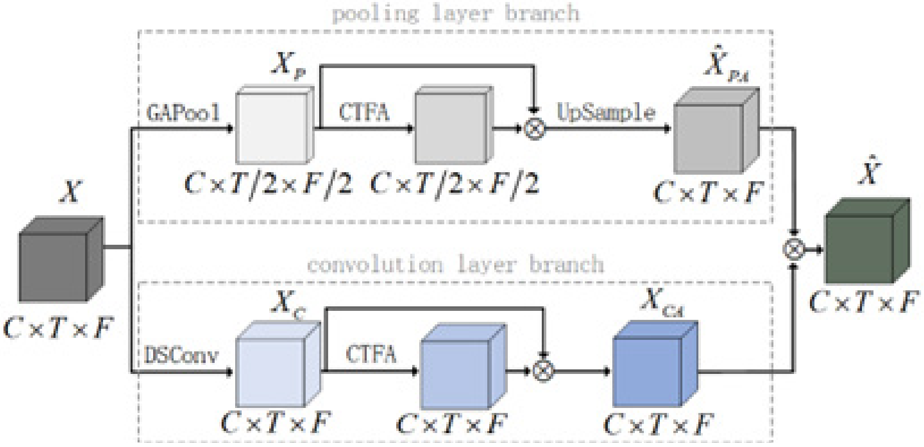


Figure 2: The structure of DCCM.

CTFA consists of Time-Frequency Fusion Module (TFFM) and Local Feature Extraction Module (LFEM) to maximize the number of reasonable feature representations and effective spoofing cues from the channel dimension, time dimension and frequency dimension. The structure is shown in Figure 3. CTFA uses the module's global information search capability to follow the interaction between the time and frequency domains to improve the system's detection performance. The module captures both channel information and maximizes the extraction of time-frequency interaction features to capture potential differentiating cues between the time and frequency dimensions, allowing the model to focus on the most feature-rich time periods and sub-bands, enhancing the feature representation of the object of interest. The module embeds information in the temporal and frequency dimensions into the channel dimension to perform adaptive feature refinement on the input features. Since the key features in spoofed speech systems are spoofing artefacts left behind after data forgery, which may not contain semantic information but rather some fine-grained feature information, LFEM extracts local fine-grained features to help the network capture more detailed information to prevent overlooking fine-grained artefacts.

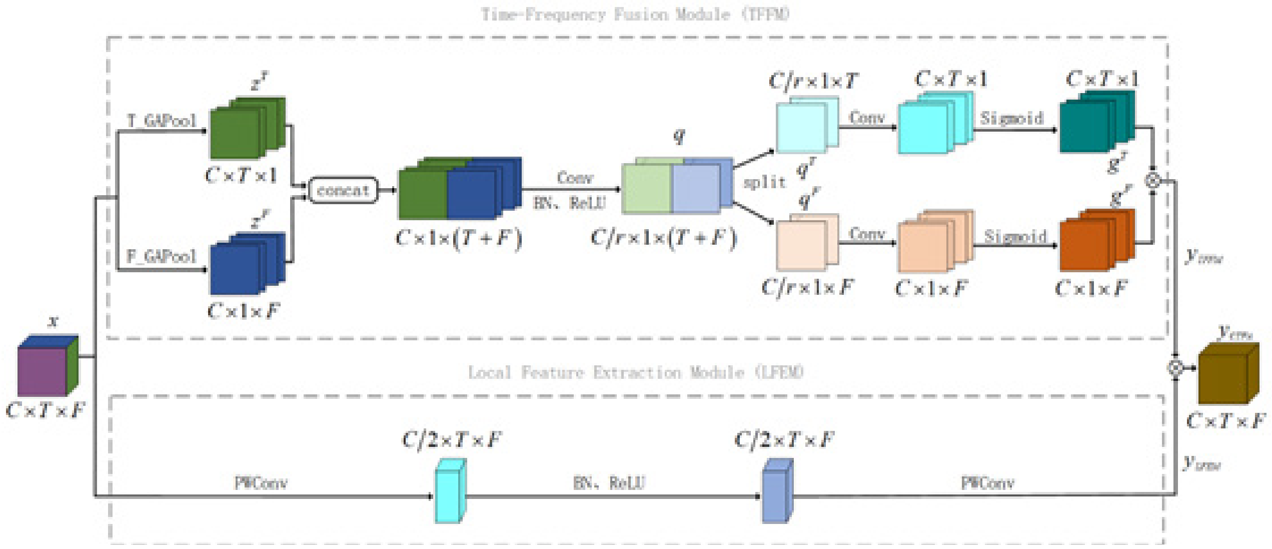


Figure 3: The structure of CTFA.

DCCM is used to extract valuable contextual information to obtain correlation information between different spoofing artifacts, fuse cross-dimensional interaction features based on attention mechanisms to aggregate distinguishing cues, and integrate cross-dimensional interaction features with contextual information to refine vital information about spoofing artifacts in order to obtain a comprehensive information feature representation, which helps improve the network's anti-interference capability and efficiently detect spoofing artifacts. CTFA is designed to capture and fuse interaction features between the time and frequency domains as well as local fine-grained features to maximize the potential for capturing discriminative cues and exploit more fine-grained feature information to prevent subtle artifacts from being overlooked.

Experimental results show that our proposed detection system has good practicality and generalizability, but its detection performance needs to be improved in the face of unknown spoofing attacks. We consider introducing data augmentation to improve the robustness of unknowns for interfering changes and maximize the capture of valid information. At the same time, we will improve the backbone network to enhance the feature extraction capability of the network and improve the generalization performance of the detection model.

References

1. Kinnunen T and Li H (2010) An overview of text-independent speaker recognition: from features to supervectors. Speech Communication 52(1): 12-40.
2. Mittal A, Dua M (2021) Automatic speaker verification systems and spoof detection techniques: Review and analysis. International Journal of Speech Technology 25: 105-134.
3. Tak H, Patino J, Todisco M (2021) End-to-end anti-spoofing with RawNet. Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE Press, Piscataway, New Jersey, USA, pp. 6369-6373.
4. Ge W Y, Patino J, Todisco M (2021) Raw differentiable architecture search for speech deepfake and spoofing detection.
5. Kang W H, Alam J, Fathan A (2022) Attentive activation function for improving end-to-end spoofing countermeasure systems.
6. Wang X, Yamagishi J, Todisco M, H Delgado, A Nautsch, et al. (2020) ASVspoof 2019: a large-scale public data base of synthesized, converted and replayed speech. Computer Speech & Language 64: 101-114.
7. Yamagishi J, Wang X, Todisco M, Sahidullah M, Patino J, et al. (2021) ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. International Speech Communication Association, pp. 47-54.