


Article

Replay Speech Detection Based on Dual-Input Hierarchical Fusion Network

Chenlei Hu ¹, Ruohua Zhou ^{1,*}  and Qingsheng Yuan ²

¹ School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 102627, China

² National Computer Network Emergency Response Technical Team Coordination Center of China, Beijing 100029, China

* Correspondence: zhouruohua@bucea.edu.cn

Abstract: Speech anti-spoofing is a crucial aspect of speaker recognition systems and has received a great deal of attention in recent years. Deep neural networks have achieved satisfactory results in datasets with similar training and testing data distributions, but their generalization ability is limited in datasets with different distributions. In this paper, we proposed a novel dual-input hierarchical fusion network (HFN) to improve the generalization ability of our model. The network had two inputs (the original speech signal and the time-reversed signal), which increased the volume and diversity of the training data. The hierarchical fusion model (HFM) enabled more thorough fusion of information from different input levels and improved model performance by fusing the two inputs after speech feature extraction. We finally evaluated the results using the ASVspoof 2021 PA (Physical Access) dataset, and the proposed system achieved an Equal Error Rate (EER) of 24.46% and a minimum tandem Detection Cost Function (min t-DCF) of 0.6708 in the test set. Compared with the four baseline systems in the ASVspoof 2021 competition, the proposed system min t-DCF values were decreased by 28.9%, 31.0%, 32.6%, and 32.9%, and the EERs were decreased by 35.7%, 38.1%, 45.4%, and 49.7%, respectively.

Keywords: anti-spoofing; replay speech detection; HFM; ASVspoof 2021



Citation: Hu, C.; Zhou, R.; Yuan, Q. Replay Speech Detection Based on Dual-Input Hierarchical Fusion Network. *Appl. Sci.* **2023**, *13*, 5350. <https://doi.org/10.3390/app13095350>

Academic Editor: Luis Javier García Villalba

Received: 13 March 2023

Revised: 19 April 2023

Accepted: 20 April 2023

Published: 25 April 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Automatic Speaker Verification (ASV) systems are crucial for biometric authentication, using acoustic features to verify that the input speech is from the target speaker [1]. However, the security of ASV systems is threatened by spoofing attacks, where attackers can create fake speech to pass as real users using methods such as replay [2], speech synthesis [3,4], and speech conversion [5,6]. As a result, both industry and academia are paying more attention to the problem of voice anti-spoofing, which aims to distinguish between spoofing attacks and real users.

ASVspoof 2021 was the fourth biannual ASVspoof challenge, and aimed to promote the development of speech anti-spoofing systems [7]. This competition differed from the previous one in that the ASVspoof 2019 PA task focused on a simulated environment where fake audio was generated. ASVspoof 2021, on the other hand, was designed to develop more powerful spoofing detection systems in more realistic environments. There are two types of fake speech generation: one is real speech recorded and played by various electronic devices in an acoustic environment, and the other is generated in a simulated environment. According to the evaluation plan in [8], model training could only use the training and development sets of ASVspoof 2019, while the test set comprised data recorded in a real environment under ASVspoof 2021. This resulted in a mismatch between the training and test data, which presented new challenges for designing speech anti-spoofing systems, requiring researchers to develop replay speech recognition systems adapted to real environments from simulated speech corpora.

An analysis of the related models of ASVspoof 2017 and 2019 revealed that the most advanced spoofing detection systems are primarily founded upon the use of convolutional neural networks (CNNs). The differences among these systems mainly focus on front-end speech feature extraction [9,10], back-end classification network architecture [11–13], and training strategies such as the loss function, learning rate scheduler, optimizer, etc. [14].

Lavrentyeva et al. [15] reported the best results in the ASVspoof 2017 competition by utilizing features such as spectrograms and identification vectors (I-Vectors) [16] as inputs, along with Light CNN (LCNN) and Recurrent Neural Network (RNN) classifiers. Huang et al. [17] proposed an attention-enhanced DenseNet-BiLSTM network model, which employs segmented linear filter group features. Their findings show that the model based on segmented linear filter group features, along with attention-enhanced DenseNet-BiLSTM, can more effectively distinguish between replayed speech. The model achieved a relative EER reduction of 74.04% on the ASVspoof 2017 dataset compared to the baseline system, which was based on constant Q cepstral coefficients and Gaussian mixture models. Todisco et al. [18] found that using multiple classifiers can enable the detection of text-to-speech synthesis attacks, voice conversion attacks, and replayed speech threats. Zeinali et al. [19] addressed the physical access problem of ASVspoof 2019 and proposed fusing Power Spectrum and Constant Q Transform (CQT) features, and trained two VGG networks. Monteiro et al. [20] proposed a deep residual network (ResNet) with self-attention for speech anti-spoofing systems, which demonstrated powerful performance in identifying spoofing cues. Lavrentyeva et al. [21] used a variant of the convolutional neural network, Light CNN (LCNN), based on Max-Feature-Map activation (MFM), which performed excellently in speech anti-spoofing detection. Zhang et al. [22] investigated deep learning models for speech anti-spoofing detection and demonstrated that a combination of CNNs and Recurrent Neural Networks (RNNs) can improve the robustness of the system. Li et al. [23] presented a technique that incorporates the Res2Net network for detecting replay and synthetic speech. The method utilized the Res2Net network to partition the feature maps into multiple channel groups, and developed a residual-like connection between these groups, thereby expanding the receptive field and generating multiple feature scales. This approach effectively enhances the detection capacity of the anti-spoofing system against unobserved spoofing attacks. Tak et al. [24] were the first to employ the RawNet2 network model in their anti-spoofing system. The method utilized RawNet2 to input raw audio, allowing it to detect some fake speech that traditional models may not. The system exhibited outstanding performance, and became the official baseline system for the ASVspoof 2021 challenge. Yoon et al. [25] introduced a multi-point input method based on convolutional neural networks for detecting replay speech attacks, which enhances the amount of information gained from a single model input. Additionally, the method was applied as a baseline for the ASVspoof 2019 PA track. Lai et al. [26] demonstrated that using different front-end feature extractions, training the same back-end residual neural network, and model fusion can lead to better recognition results. These models showed strong capabilities in time domain and frequency domain modeling, and exhibited excellent performance on the ASVspoof 2019 dataset.

Due to the increased focus in the ASVspoof 2021 PA test set on false audio in real-world scenarios, traditional DNN models cannot achieve outstanding performance on the new track test set. The reason for this is that the generalization ability of traditional DNN models is limited in cross-domain scenarios, resulting in poor adaptation to situations with different training and testing distributions. The HFN model in the field of image anti-spoofing extracts information from different feature levels by fusing the original RGB image information and the corresponding meta-pattern information, thus effectively improving the generalization capability of face anti-fraud systems [27]. Similarly, in the field of speech recognition, using both the original speech signal and its time-reversed version as inputs can improve the robustness of the model [28]. This is because the time-reversed signal and the original signal have different features in the frequency domain, which can help the classification network adapt to different speech environments [29].

To improve the generalization ability of replayed speech anti-spoofing models, this paper introduces the HFN model architecture [27] into the replayed speech anti-spoofing system, and proposes using both original speech signals and time-reversed speech signals as inputs to the network. Advanced performance was achieved in the ASVspoof 2021 PA task. The main contributions of our work are as follows:

1. We first introduce the HFN model into the replay attack detection system. The HFM in the model can more thoroughly fuse information from inputs at different levels, thereby improving the robustness of the system.
2. We propose using both the original speech signal and the time-reversed speech signal as inputs to the network, which increases the amount of information provided to the model and facilitates the learning and classification of the model.
3. Our proposed method achieves an EER of 24.46% and a min t-DCF of 0.6708 on the ASVspoof 2021 PA test set. Compared to the four baseline systems of the ASVspoof 2021 PA task, the min t-DCF values are reduced by 28.9%, 31.0%, 32.6%, and 32.9%, and the EER values are reduced by 35.7%, 38.1%, 45.4%, and 49.7%, respectively.

The remainder of the paper is as follows. Section 2 illustrates our method of HFN and the input methods. Section 3 describes the dataset and evaluation indicators used in this study. Section 4 presents the experimental results and a comparison with other methods. In Section 5, we conclude the paper.

2. Methods

2.1. Hierarchical Fusion Network

The proposed network had two inputs, as shown in Figure 1a, namely, the original speech signal and the time-reversal speech signal. These two signals came from the same speech segment, with the time-reversed speech signal inverted on the time axis relative to the original speech signal. To integrate the original speech signals and the associated time-reversal signals, we used a dual-input hierarchical fusion network (HFN), as shown in Figure 1b. The top network architecture of HFN was for processing the original signals, and the bottom network architecture was for processing the time-reversed signals. The model architectures of the original and reversed layers were the same. For the backbone part of the model, we used ResNet-50 [30] in each layer. Features from the ends of the two layers were merged with the fully connected layer to produce a binary vector for classification. In addition, our model improved fusion performance by more thoroughly fusing information from different feature levels, as shown in Figure 1c, where features from different feature levels were progressively fused by the hierarchical fusion model (HFM) [27].

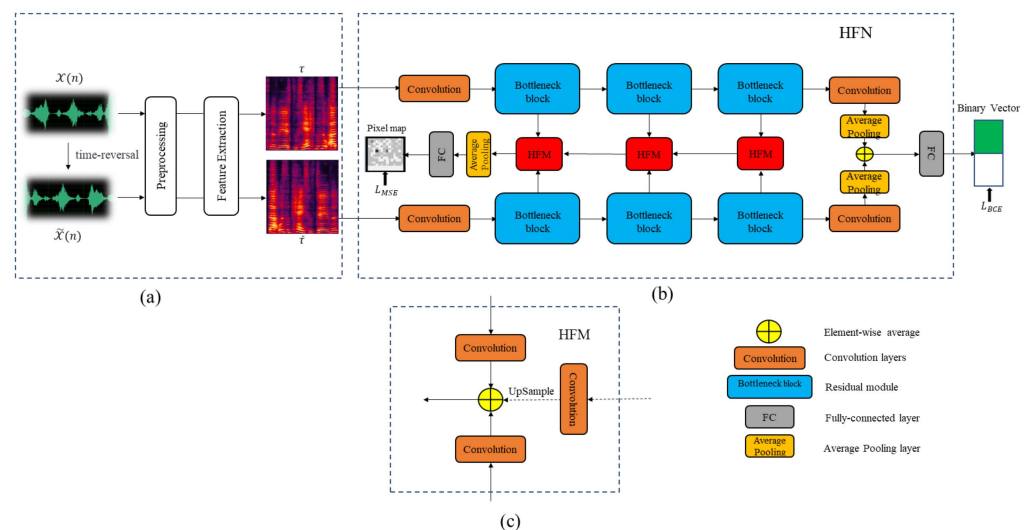


Figure 1. Diagram of dual-input hierarchical fusion network. (a) Original speech and time-flipped speech signal input; (b) hierarchical fusion network (HFN); (c) hierarchical fusion model (HFM).

In the HFN network, the F_j of the j th HFM module combines feature maps from the top-level raw stream z_j^t , the reversed stream z_j^b , and the fused result from the previous hierarchy z_{j-1} . z_j can be expressed as:

$$z_j = F_j(z_j^t, z_j^b, z_{j-1}) = \varphi(z_j^t) + \varphi(z_j^b) + \theta[\varphi(z_{j-1})], j = 1, 2, 3, z_0 = 0 \quad (1)$$

where φ is a convolutional layer used to align the number of channels of feature maps from different hierarchies, and θ is the closest interpolation function used to upsample z_{j-1} so that it has the same size as both $\varphi(z_j^b)$ and $\varphi(z_j^t)$. The information fusion from multiple feature levels was utilized to address the face anti-spoofing issue [27].

In order to anticipate a pixel map for pixel-level supervision, a fused feature map corresponding to a binary label was constructed at the final fusion level [31,32]. The fused feature map was refined during the decoding process using pixel-level supervision, which optimizes the HFN by combining binary classification with pixel-wise supervision. In our experiments, the binary cross entropy loss (L_{BCE}) was applied to the binary vector output from the fully connected layer, where the fake speech was marked as "0" and genuine speech was marked as "1". The output pixel map of the HFN was then subjected to mean squared error (MSE) loss (L_{MSE}). Each fake or genuine speech sample was given a target pixel map with all components set to "0" or "1", respectively, corresponding to the binary labels. Thus, the final loss function of the model was:

$$L_{loss} = L_{BCE} + L_{MSE} \quad (2)$$

2.2. Different Input Methods of Models

In this paper we involve three different model input approaches: the single-input method, the signal overlay-input method, and the dual-input method. To validate the effectiveness of our proposed methods, we designed comparative experiments with ResNet50 as the selected classification network. This is because the model structure of the original and inverted layers of the hierarchical fusion network is based on ResNet50.

For the ResNet50-based model, we used both single-input and signal overlay-input approaches, which involved: (1) using the original speech signal and the time-reversed speech signal as inputs to the classification network after feature extraction, and (2) overlaying the original speech signal and time-reversed speech signal as inputs to the classification network [33].

For the dual-input Hierarchical Fusion Model proposed in this study, the model input was dual-input, and three strategies were employed in the experiment: (1) both the original speech signal and time-reversed speech were used as inputs; (2) the original speech signal was used as an input for both the upper and lower layers of the HFN; and (3) the time-reversed speech signal was used as an input for both the upper and lower layers of the HFN.

3. Research Process

3.1. Datasets and Evaluation Metrics

The data distribution of the training set, development set, and evaluation set for the PA task of the ASVspoof 2021 challenge is shown in Table 1. The training set provided in the ASVspoof 2019 PA dataset consisted of 5400 genuine and 48,600 spoofed utterances, while the development set consisted of 5400 genuine and 24,300 spoofed utterances. The final evaluation set was based on the ASVspoof 2021 PA dataset. In this paper, the training set was used to train the replay speech spoofing detection model, and the evaluation set was utilized to submit the final results.

Table 1. The distribution of ASVspoof dataset.

Dataset	Bona Fide	Spoofer
	Utterance	Utterance
Train	5400	48,600
Dev	5400	24,300
Evaluation	94,068	627,264

The ASVspoof challenges use two evaluation metrics [8]. The primary evaluation metric for the PA task is the Minimum Tandem Detection Cost Function (min t-DCF) [34], which is defined as in Equation (3):

$$t - DCF(\theta) = C_1 P_{miss}^{cm}(\theta) + C_2 P_{fa}^{cm}(\theta) \quad (3)$$

In this equation, $P_{miss}^{cm}(\theta)$ and $P_{fa}^{cm}(\theta)$ represent the false rejection rate and false acceptance rate of the speech replay detection system, respectively; θ is the threshold of the model. Additionally, the min t-DCF metric does not require a preset threshold.

The secondary metric for the PA task is the EER, which is defined as follows:

$$P_{fa}(\theta) = \frac{\#\{\text{fake trials with score} > \theta\}}{\#\{\text{total fake trials}\}} \quad (4)$$

$$P_{miss}(\theta) = \frac{\#\{\text{genuine trials with score} < \theta\}}{\#\{\text{total genuine trials}\}} \quad (5)$$

where $P_{fa}(\theta)$ and $P_{miss}(\theta)$ represent the false acceptance rate and false rejection rate at threshold θ , respectively. Additionally, the EER corresponds to the threshold θ_{EER} at which the false acceptance rate and false rejection rate are equal. In addition, the ASVspoof competition also sets up a parameter-fixed ASV system to help researchers to focus on the design of the replay speech detection model.

3.2. Feature Extraction

To verify the effectiveness of the proposed model, three acoustic features were used for experimental evaluation in this paper, including a log power magnitude spectrogram (Spec), Linear Frequency Cepstral Coefficients (LFCC), and Constant Q Transform (CQT).

The log power magnitude spectrogram, calculated using short-time Fourier transform (STFT) [35], is a 3D graph that represents time, frequency, and signal strength. The speech signal is divided into several short-time windows of equal duration, and the frequency components of each window are computed. Spec is obtained by taking the logarithm and square of the STFT output, resulting in a more intuitive form of spectral analysis. This feature better simulates the nonlinear characteristics of human auditory perception, which is nonlinear in terms of sound intensity and frequency perception. Furthermore, the logarithmic and squared output can amplify low-frequency signals and compress high-frequency signals, leading to enhanced visualization of low-frequency signals. In various fields, such as speech recognition, Spec is widely utilized as an input feature to extract spectral information from speech signals, facilitating the recognition of speech content by machine learning and deep learning algorithms.

Linear Frequency Cepstral Coefficients are a powerful method used to identify and classify amplitude characteristics in different speech signals [36]. This method is an improvement over traditional Mel Frequency Cepstral Coefficients (MFCC), as the filters in LFCC are uniformly distributed across the frequency range, allowing for the comprehensive capture of speech signal information. On the other hand, the filter bank in MFCC is only designed based on the human ear's sensitivity to sounds with different frequencies, resulting in a lower resolution in the high-frequency region. In the calculation process of LFCC, the output of each filter is logarithmically transformed into a power spectrum

and further converted into cepstral coefficients using discrete cosine transform. Due to the high correlation between these coefficients, only the first few are required to represent speech signal features. The cepstral coefficients obtained by employing this method are called static LFCC coefficients. Moreover, dynamic LFCC coefficients can be obtained by computing first- and second-order differences in the cepstral coefficients, providing more information about instantaneous changes in the speech signal. Therefore, LFCC is an effective method for feature extraction in speech signals and has a significant advantage in processing high-frequency information in synthetic speech; moreover, the combined use of static and dynamic LFCC coefficients is common in representing speech signal features.

Constant Q Transform is a classic time–frequency transform used to analyze the spectral characteristics of a signal [37]. It differs from traditional time–frequency analysis methods, such as Fourier transform, in that it filters the time-domain speech signal with a set of constant Q filters. This is because the filter satisfies the constant Q constraint, i.e., the center frequency-to-bandwidth ratio is the same and the bandwidth magnitude is different at low and high frequencies, resulting in a non-linear frequency domain signal. CQT avoids the disadvantage of uniform time–frequency resolution; low-frequency waves have a narrower bandwidth, and therefore, higher frequency resolution to decompose similar notes, and high-frequency waves have a broader bandwidth, and therefore, higher temporal resolution to track rapidly changing overtones. Compared to traditional Fourier transform, it is more flexible in balancing temporal and frequency resolution, allowing for more accurate spectral analysis of complex speech signals while reducing computational complexity, and is therefore widely used in music and speech signal processing.

The details of the feature extraction parameters are as follows: (1) For Spec: a Hanning window with a size of 25 ms, a step of 10 ms, and 512 Fast Fourier Transform (FFT) points; (2) For LFCC: a 60-dimensional feature vector, a window length of 20 ms, 512 FFT points, 20 filters, plus delta and double delta coefficients; (3) For CQT: a Hanning window, nine octaves, 48 bins per octave, and a step of 16 ms. All features were trimmed along the temporal axis to maintain an accurate 400 frames [23].

3.3. Training Details

The proposed model used binary cross-entropy and the mean squared error as loss functions. In addition, Adam [38] was used as the optimizer, where $\beta_1 = 0.9$, $\beta_2 = 0.85$, and weight decay is set to 10^{-9} . After training for 30 epochs, the model with the lowest EER in the development set was selected for evaluation.

3.4. Test Details

During the test phase, the score obtained by combining the output binary vector and the pixel map was used for classification, where $S = [S_0, S_1]$ is the output binary vector, with $S_0, S_1 \in [0, 1]$ and $S_0 + S_1 = 1$, and $m \in \{R^{+32 \times 32} | m(i, j) \in [0, 1]\}$ corresponds to the output of the pixel map. The score $S \in [0, 1]$ for classification was calculated using Equation (6):

$$S = \frac{S_1 + \text{mean}(m)}{2} \quad (6)$$

where the score $S \in [0, 1]$ represents the probability that the input speech is “genuine”.

4. Results

4.1. The System Performance of Different Input Methods

In this section, we evaluate the effectiveness of the proposed replay speech detection system based on a dual-input hierarchical fusion approach. In order to verify the effectiveness of the proposed method, ResNet50 and HFN were used as classification networks in this paper. At the same time, different input methods were employed, including the single-input, overlay-input, and dual-input methods. Furthermore, CQT was chosen as the acoustic feature for the model input. As shown in Table 2, when applying the original speech signal and time-reversed speech signal as inputs to the ResNet50 and HFN

models, the min t-DCF is better than for other input methods. Furthermore, our proposed dual-input hierarchical fusion network exhibits significantly better performance than the ResNet50 model. The experimental results show that the best system has an EER of 28.51% in the test set of the PA track. Compared to the best system that employs the ResNet50 network, the proposed system achieves a relative reduction in EER of 25.3%.

Table 2. The EER (%) and min t-DCF of the proposed system corresponding to different input methods.

Input Methods	Acoustic Features	Models	Min t-DCF	EER (%)
Single-input	CQT	ResNet-50	0.9913	38.17
	CQT (invert)		0.9848	41.12
Overlay-input	CQT + CQT (invert)	ResNet-50	0.9826	38.84
Dual-input	CQT + CQT (invert)	HFN	0.7572	28.51
	Double CQT		0.7741	29.33
	Double CQT (invert)		0.7709	29.63

4.2. The System Performance of Different Acoustic Features

This section evaluates the system performance of the dual-input hierarchical fusion network using different acoustic features of the ASVspoof 2021 PA test set, and compares it with that of the baseline system. Table 3 shows the min t-DCF and EER of the baseline system and our proposed system in the evaluation test. The experimental results show that using CQT as the model input feature achieved the best system performance. CQT is a classic time–frequency transform method with advantages such as higher frequency resolution, balanced time–frequency resolution, and lower computational complexity, and is widely used in fields such as music signal classification and speech anti-spoofing. Similarly, using LFCC as the model input feature also achieved desirable results. LFCC is the official baseline feature of ASVspoof and is an amplitude feature extraction method. Compared with traditional acoustic features, such as Mel-frequency cepstral coefficients, its filter bank is uniformly distributed over the entire frequency range; thus, it can more comprehensively capture information in speech signals. In addition, using Spec as the model input feature achieved dissatisfactory results. This is because the Spec feature is the logarithmic amplitude spectrum and does not contain phase information. However, the synthetic speech in real-world environments may exist in the phase information of speech, so the system performance of the Spec feature is relatively inferior to the other two acoustic features.

Table 3. The EER (%) and min t-DCF of the proposed system corresponding to different acoustic features.

System	Min t-DCF	EER (%)	
Baseline	CQCC + GMM	0.9434	38.07
	LFCC + GMM	0.9724	39.54
	LFCC + LCNN	0.9958	44.77
	RawNet2	0.9997	48.60
Proposed	HFN + Spec	0.9453	45.15
	HFN + LFCC	0.8945	36.25
	HFN + CQT	0.7572	28.51
	Fusion	0.6708	24.46

According to the results in Table 3, the HFN + CQT and HFN + LFCC models we proposed perform significantly better than all the baseline systems provided by the ASVspoof 2021 PA task. Among them, the HFN + LFCC model reduces the min t-DCF and EER relative to the baseline system by a maximum of 10.5% and 25.4%, respectively; the HFN + CQT model reduces the min t-DCF and EER relative to the baseline system

by a maximum of 24.3% and 41.3%, respectively. The proposed fusion system is the weighted average of the scores of the three single systems: HFN + Spec, HFN + LFCC, and HFN + CQT. The fusion system achieves the best results in the final evaluation set, with an EER of 24.46% and a min t-DCF of 0.6708. Compared with the four baseline systems in the ASVspoof 2021 competition, in our fusion model, the min t-DCF values are decreased by 28.9%, 31.0%, 32.6%, and 32.9%, and the EER values are reduced by 35.7%, 38.1%, 45.4%, and 49.7%, respectively.

4.3. The System Performance Compared with State-of-the-Art Systems

Table 4 presents the experimental results of the top five teams in the ASVspoof 2021 Challenge [7]. As can be seen from the results of the participating teams, only the top result had a min t-DCF below 0.7 and an EER below 25% in the PA evaluation set, demonstrating that developing a replay speech detection system from a simulated speech corpus that perfectly matches the real environment is a significant challenge.

Among them, T07, a team from Kunshan Duke University [39], achieved the best performance in the PA track. They proposed using the difference between the codec-filtered audio and the original audio as input to the model, using codecs such as WORLD [40] and MelGAN [41]; they also used speech enhancement methods such as velocity perturbation and reverberation to augment the model training data; in addition, several outlier detection systems were used as classifiers, with the final submission being an average of various systems. T23, a team from Russia, proposed using a variety of front ends, including Mel-scaled short-time Fourier transform of an audio signal's linear or rectangular frequency filter banks, and trainable transforms such as LEAF or SinConv [24]; moreover, several deep neural networks such as ResNet18, LCNN9, RawNet2, and their modifications, were used as classifiers, and microphone and room impulse responses were used for data enhancement. The final system is a fusion of several classification models [42].

Table 4. Comparison with the experimental results of the top five participating teams in the ASVspoof 2021 challenge.

Teams	Min t-DCF	EER (%)
T07 [39]	0.6824	24.25
T16	0.7122	27.59
T23 [42]	0.7296	26.42
T01	0.7446	28.36
T04	0.7462	29.00
Proposed method	0.6708	24.46

Instead of using speech enhancement, our proposed method takes both the original audio signal and the time-reversed audio signal as inputs to the model and uses HFN as the classifier, with the fusion system being a weighted average of the three single system scores of HFN + Spec, HFN + LFCC, and HFN + CQT. Compared with the best results, the proposed system improved the min t-DCF by 1.7% and achieved a similar level of EER.

5. Conclusions

This work proposes a replay speech detection method based on a dual-input hierarchical fusion network. First, the model was designed with two inputs, namely, the original speech signal and the time-reversed speech signal, to increase the information content of the model input. Then, a classification network model was constructed, and the HFN architecture was introduced into the ASV anti-spoofing system for the first time. The HFN contained an HFM fusion module that can fuse the output of different residual blocks of the upper and lower network layers, effectively addressing the domain discrepancy problem between the training and test sets and improving the model's generality. The results show that our proposed network outperforms the baseline system provided in the ASVspoof 2021 PA task, as well as the systems of other participating teams, in min t-DCF.

Author Contributions: Conceptualization, C.H., R.Z. and Q.Y.; methodology, C.H.; software, C.H.; validation, C.H. and R.Z.; formal analysis, C.H.; investigation, C.H.; resources, R.Z.; data curation, R.Z.; writing—original draft preparation, C.H.; writing—review and editing, C.H., R.Z. and Q.Y.; visu-alization, Q.Y.; supervision, R.Z. and Q.Y.; project administration, R.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The ASVspooF 2021 PA dataset used in this paper is available through the following link: (<https://zenodo.org/record/4834716> accessed on 13 March 2023).

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

References

1. Delac, K.; Grgic, M. A survey of biometric recognition methods. In Proceedings of the Elmar-2004, 46th International Symposium on Electronics in Marine, Zadar, Croatia, 18 June 2004; IEEE: New York, NY, USA; pp. 184–193.
2. Kinnunen, T.; Evans, N.; Yamagishi, J.; Lee, K.A.; Sahidullah, M.; Todisco, M.; Delgado, H. AsvspooF 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *Training* **2017**, *10*, 1508.
3. Juvela, L.; Bollepalli, B.; Wang, X.; Kameoka, H.; Airaksinen, M.; Yamagishi, J.; Alku, P. Speech waveform synthesis from MFCC sequences with generative adversarial networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 1–20 April 2018; IEEE: New York, NY, USA; pp. 5679–5683.
4. Oord, A.V.D.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
5. Toda, T.; Chen, L.H.; Saito, D.; Villavicencio, F.; Wester, M.; Wu, Z.; Yamagishi, J. The Voice Conversion Challenge 2016. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016; pp. 1632–1636.
6. Huang, W.C.; Lo, C.C.; Hwang, H.T.; Tsao, Y.; Wang, H.M. Wavenet vocoder and its applications in voice conversion. In Proceedings of the 30th ROCLING Conference on Computational Linguistics and Speech Processing (ROCLING), Hsinchu, Taiwan, 4–5 October 2018.
7. Yamagishi, J.; Wang, X.; Todisco, M.; Sahidullah, M.; Patino, J.; Nautsch, A.; Liu, X.; Lee, K.A.; Kinnunen, T.; Delgado, H.; et al. ASVspooF 2021: Accelerating progress in spoofed and deepfake speech detection. *arXiv* **2021**, arXiv:2109.00537.
8. Delgado, H.; Evans, N.; Kinnunen, T.; Lee, K.A.; Liu, X.; Nautsch, A.; Patino, J.; Sahidullah, M.; Todisco, M.; Yamagishi, J.; et al. AsvspooF 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *arXiv* **2021**, arXiv:2109.00535.
9. Li, X.; Li, N.; Zhong, J.; Wu, X.; Liu, X.; Su, D.; Yu, D.; Meng, H. Investigating robustness of adversarial samples detection for automatic speaker verification. *arXiv* **2020**, arXiv:2006.06186.
10. Cheng, X.; Xu, M.; Zheng, T.F. Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspooF 2019. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 540–545.
11. Wu, Z.; Chng, E.S.; Li, H. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012.
12. Alzantot, M.; Wang, Z.; Srivastava, M.B. Deep residual neural networks for audio spoofing detection. *arXiv* **2019**, arXiv:1907.00501.
13. Cai, W.; Wu, H.; Cai, D.; Li, M. The DKU replay detection system for the ASVspooF 2019 challenge: On data augmentation, feature representation, classification, and fusion. *arXiv* **2019**, arXiv:1907.02663.
14. Zhang, Y.; Jiang, F.; Duan, Z. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Process. Lett.* **2021**, *28*, 937–941. [[CrossRef](#)]
15. Lavrentyeva, G.; Novoselov, S.; Malykh, E.; Kozlov, A.; Kudashev, O.; Shchemelinin, V. Audio Replay Attack Detection with Deep Learning Frameworks. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 82–86.
16. Garcia-Romero, D.; Espy-Wilson, C.Y. Analysis of i-vector length normalization in speaker recognition systems. In Proceedings of the Twelfth Annual Conference of the International Speech Communication Association, Florence, Italy, 27–31 August 2011.
17. Huang, L.; Pun, C.M. Audio replay spoof attack detection by joint segment-based linear filter bank feature extraction and attention-enhanced DenseNet-BiLSTM network. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1813–1825. [[CrossRef](#)]
18. Todisco, M.; Wang, X.; Vestman, V.; Sahidullah, M.; Delgado, H.; Nautsch, A.; Yamagishi, J.; Evans, N.; Kinnunen, T.; Lee, K.A.; et al. ASVspooF 2019: Future horizons in spoofed and fake audio detection. *arXiv* **2019**, arXiv:1904.05441.
19. Zeinali, H.; Stafylakis, T.; Athanasopoulou, G.; Rohdin, J.; Gkinis, I.; Burget, L.; Černocký, J. Detecting spoofing attacks using vgg and sincnet: But-omilia submission to asvspooF 2019 challenge. *arXiv* **2019**, arXiv:1907.12908.

20. Monteiro, J.; Alam, J.; Falk, T.H. Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers. *Comput. Speech Lang.* **2020**, *63*, 101096. [[CrossRef](#)]
21. Lavrentyeva, G.; Novoselov, S.; Tseren, A.; Volkova, M.; Gorlanov, A.; Kozlov, A. STC antispoofing systems for the ASVspoof2019 challenge. *arXiv* **2019**, arXiv:1904.05576.
22. Zhang, C.; Yu, C.; Hansen, J.H.L. An Investigation of Deep-Learning Frameworks for Speaker Verification Antispoofing. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 684–694. [[CrossRef](#)]
23. Li, X.; Li, N.; Weng, C.; Liu, X.; Su, D.; Yu, D.; Meng, H. Replay and synthetic speech detection with res2net architecture. 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. 6354–6358.
24. 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.
25. Yoon, S.H.; Yu, H.J. Multiple points input for convolutional neural networks in replay attack detection. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6444–6448.
26. Lai, C.I.; Chen, N.; Villalba, J.; Dehak, N. ASSERT: Anti-spoofing with squeeze-excitation and residual networks. *arXiv* **2019**, arXiv:1904.01120.
27. Cai, R.; Li, Z.; Wan, R.; Li, H.; Hu, Y.; Kot, A.C. Learning meta pattern for face anti-spoofing. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 1201–1213. [[CrossRef](#)]
28. Chao, F.A.; Jiang SW, F.; Yan, B.C.; Hung, J.W.; Chen, B. TENET: A time-reversal enhancement network for noise-robust ASR. In Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 13–17 December 2021; pp. 55–61.
29. Yoon, S.H.; Koh, M.S.; Yu, H.J. Phase Spectrum of Time-Flipped Speech Signals for Robust Spoofing Detection. In Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2020), Tokyo, Japan, 1–5 November 2020; pp. 319–325.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
31. Sun, W.; Song, Y.; Chen, C.; Huang, J.; Kot, A.C. Face spoofing detection based on local ternary label supervision in fully convolutional networks. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 3181–3196. [[CrossRef](#)]
32. Chen, B.; Yang, W.; Li, H.; Wang, S.; Kwong, S. Camera Invariant Feature Learning for Generalized Face Anti-Spoofing. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 2477–2492. [[CrossRef](#)]
33. Yoon, S.; Yu, H.J. Multiple-point input and time-inverted speech signal for the ASVspoof 2021 challenge. In Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, Online, 16 September 2021; pp. 37–41.
34. Kinnunen, T.; Lee, K.A.; Delgado, H.; Evans, N.; Todisco, M.; Sahidullah, M.; Yamagishi, J.; Reynolds, D.A. t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification. *arXiv* **2018**, arXiv:1804.09618.
35. Griffin, D.; Lim, J. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 236–243. [[CrossRef](#)]
36. Hueber, T.; Chollet, G.; Denby, B.; Stone, M. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. *Proc. ISSP* **2008**, *24*, 365–369.
37. Brown, J.C. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *J. Acoust. Soc. Am.* **1999**, *105*, 1933–1941. [[CrossRef](#)] [[PubMed](#)]
38. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
39. Wang, X.; Qin, X.; Zhu, T.; Wang, C.; Zhang, S.; Li, M. The DKU-CMRI system for the ASVspoof 2021 challenge: Vocoder based replay channel response estimation. In Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, Online, 16 September 2021; pp. 16–21.
40. 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*, 1877–1884. [[CrossRef](#)]
41. Kumar, K.; Kumar, R.; de Boissiere, T.; Gestin, L.; Teoh, W.Z.; Sotelo, J.; de Brébisson, A.; Bengio, Y.; Courville, A.C. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
42. Tomilov, A.; Svishchev, A.; Volkova, M.; Chirkovskiy, A.; Kondratev, A.; Lavrentyeva, G. STC antispoofing systems for the ASVspoof2021 challenge. In Proceedings of the ASVspoof 2021 Workshop, Online, 16 September 2021; pp. 61–67.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.