applied sciences

MDPI

Article

Voice Deepfake Detection Using the Self-Supervised Pre-Training Model HuBERT

Lanting Li, Tianliang Lu *, Xingbang Ma, Mengjiao Yuan and Da Wan

College of Information and Cyber Security, People's Public Security University of China, Beijing 100038, China
* Correspondence: lutianliang@ppsuc.edu.cn

Abstract: In recent years, voice deepfake technology has developed rapidly, but current detection methods have the problems of insufficient detection generalization and insufficient feature extraction for unknown attacks. This paper presents a forged speech detection method (HuRawNet2_modified) based on a self-supervised pre-trained model (HuBERT) to improve detection (and address the above problems). A combination of impulsive signal-dependent additive noise and additive white Gaussian noise was adopted for data boosting and augmentation, and the HuBERT model was fine-tuned on different language databases. On this basis, the size of the extracted feature maps was modified independently by the α-feature map scaling (α-FMS) method, with a modified end-to-end method using the RawNet2 model as the backbone structure. The results showed that the HuBERT model could extract features more comprehensively and accurately. The best evaluation indicators were an equal error rate (EER) of 2.89% and a minimum tandem detection cost function (min t-DCF) of 0.2182 on the database of the ASVspoof2021 LA challenge, which verified the effectiveness of the detection method proposed in this paper. Compared with the baseline systems in databases of the ASVspoof 2021 LA challenge and the FMFCC-A, the values of EER and min t-DCF decreased. The results also showed that the self-supervised pre-trained model with fine-tuning can extract acoustic features across languages. And the detection can be slightly improved when the languages of the pre-trained database, and the fine-tuned and tested database are the same.

Keywords: voice deepfake detection; self-supervised learning; pre-training; feature map scaling; anti-spoofing

Citation: Li, L.; Lu, T.; Ma, X.; Yuan, M.; Wan, D. Voice Deepfake Detection Using the Self-Supervised Pre-Training Model HuBERT. *Appl. Sci.* **2023**, *13*, 8488. https://doi.org/10.3390/app13148488

Academic Editor: David Megías

Received: 29 May 2023
Revised: 2 July 2023
Accepted: 17 July 2023
Published: 22 July 2023

1. Introduction

Audio deepfake technology has received less attention and it emerged later than face deepfake technology. At present, Baidu, Alibaba, Microsoft, Amazon, and other companies have opened speech synthesis tools to the public, which has gradually reduced the threshold and difficulty of using audio deepfake technology. In addition, the naturalness and anthropomorphic degree have been greatly improved, even to the extent that the human ear cannot distinguish between real and fake. Audio recordings have gradually become one of the most common pieces of evidence in litigation with the increased power of WeChat voice and the increasing number of portable technology products that can be used for recording. It also means that once criminals use audio deepfake technology to implement criminal activities such as fraud and the fabrication of evidence, the authenticity, integrity, and relevance of recorded materials cannot be guaranteed, which will have a terrible impact on judicial practice.

The design of the network structure, loss function, and training method can improve the performance of an audio deepfake detection model, but the potential of the model depends fundamentally on the initial features of the information captured. The production of hand-crafted features will result in a loss of some information, dramatically affecting the detection of unknown attacks. Therefore, we need more efficient and more general representations. In recent years, self-supervised learning has attracted broad concern.

Current research shows that self-supervised speech models can extract robust acoustic features for unknown domains. In addition, the existing audio deepfake detection methods are easy to overfit on the training set and have poor robustness in terms of audio detection after recompression, noise addition, and other processing. Most of the existing audio deepfake detection methods rely on specific datasets or specific deepfake methods, with a single and homogeneous distribution of training data. Most are detected on English datasets, so their generalizability cannot be tested on Chinese datasets.

To address these issues, this study proposed a self-supervised pre-trained model for brevity, namely HuRawNet2_modified. The main contributions are as follows:

(1) Regarding front-end feature extraction, self-supervised pre-training models trained on either English or Chinese datasets were used, and fine-tuning with English and Chinese datasets was undertaken to explore the impact of pre-training models using different language datasets on the results.
(2) For the back-end model, an improved end-to-end RawNet2 model was used as the backbone structure, and α-FMS was used to independently modify the size of the feature maps to improve the model detection and compare the performance with current state-of-the-art algorithms on Chinese and English datasets.
(3) In terms of datasets, to address the problem of voice deepfake detection being trained chiefly and tested on English datasets, cross-library tests were conducted on different language datasets to verify the proposed method's detection performance and generalizability on Chinese and English datasets.

2. Related Work

2.1. Detection Methods Based on Traditional Features and Related Events

Early audio deepfake detection mainly relied on hidden Markov chains and Gaussian mixture models, and later evolved into front-end and back-end models. The typical audio deepfake detection system is a framework composed of a front end and back end. The front ends extract acoustic features from speech, and the back end converts features into scores. Traditional front-end feature extractors use digital signal processing algorithms to extract spectrum, phase, or other acoustic features. Among them, the most widely used include mel-frequency cepstral coefficients (MFCC), linear frequency cepstral coefficients (LFCC), and constant-Q transform cepstral features (CQCC) [1]. However, the detection method based on traditional features will result in the loss of some information. Moreover, it is usually only effective for detecting specific types of voice deepfakes, and generalization and robustness need to be improved.

The distinguishing features of the front end of the traditional detection system adopt the hand-crafted features designed by experts, and the back end directly uses Gaussian mixture models (GMM) or support vector machine (SVM) for classification and judgment. In recent years, deep-learning-based systems have gradually become mainstream. The front ends extract the speech features of the input neural network, and the back end learns the high-level representation of the features through the neural network and then performs a classification judgment to identify the authenticity of the audio [2]. With the development of deep learning, it is increasingly common to use a deep neural network (DNN) to process the original waveform directly in many tasks. Tak et al. [3] applied the improved RawNet2 network to synthetic speech detection, used a set of sinc filters to operate the original waveform through time-domain convolution directly, and then learned deep-level discriminative information through the residual module and gate recurrent unit(GRU). Based on this network, the RawGAT-ST model was proposed [4], and the spectro-temporal graph attention network was used to model the relationship across different sub-bands and temporal segments. Based on ResNet's skip layer connection and Inception's parallel convolution structure, Hua et al. [2] designed two lightweight end-to-end time-domain synthetic speech detection networks (TSSDNet).

In order to promote the research of audio deepfake detection technology, the Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof), jointly

launched by the University of Edinburgh and other research institutions, has been held four times since 2015. This is the main event regarding audio deepfake detection [5]. The ASVspoof 2019 challenge [6] was the first challenge to consider three kinds of spoofing attacks simultaneously. The minimum tandem detection cost function (min t-DCF) was introduced to represent the performance of the whole system. The three sub-tasks of the ASVspoof 2021 challenge [7] focused on logical access(LA), physical access (PA), and speech deepfake (DF) tasks. In order to promote the development of audio deepfake detection in Chinese scenes, domestic scholars have also launched Chinese audio deepfake events, such as the second Fake Media Forensic Challenge of CSIG [8] and the third Chinese AI competition.

2.2. Detection Method Based on a Self-Supervised Speech Model

The self-supervised learning speech model is a rapidly developing research topic, and many pre-trained models have been released and used for various downstream tasks. The self-supervised speech model extracts general speech representations from speech using a self-supervised method without labels (using auxiliary information pretext) [9]. Self-supervised learning is first applied to the field of natural language processing (NLP) and computer vision (CV), saving a lot of research time and cost, and achieving good results, such as BERT [10] and word2vec [11]. However, speech signals differ from text and images, and are continuous value sequences. Therefore, self-supervised learning for audio deepfake detection faces challenges different to those associated with CV and NLP. Firstly, multiple sounds in each input statement break the instance classification assumption used in many CV domain pre-training methods. Secondly, no prior lexical dictionary of discrete sound units are available during pre-training, and it is difficult to predict the loss. Contrastive predictive coding (CPC) [12] first applied self-supervised learning to the field of automatic speech recognition (ASR) and proposed InfoNCE loss for the first time. After that, Facebook Lab proposed the classic wave2vec model [13], which is the basis of a series of models.

Despite the costs associated with training self-supervised speech models, there are a number of pre-trained self-supervised models available, such as wave2vec2.0 [14], HuBERT [9], and WavLM [15]. The most popular are the HuBERT and wave2vec 2.0 models. Previous studies have shown that in computer vision (CV) and natural language processing (NLP), self-supervised learning has apparent advantages in terms of corresponding downstream tasks. HuBERT performs pre-training on 960 h of the LibriSpeech dataset or 60,000 h of the Libri-Light dataset to obtain pre-trained speech models of various scales, which can be applied to multiple scenarios after fine-tuning. In addition, Zhang et al. [16] trained Chinese speech pre-training models of multiple scales based on the WenetSpeech dataset, which can be applied to downstream tasks in Chinese scenarios. Wang et al. [17] first introduced the self-supervised pre-trained speech model into the audio deepfake detection scene and used the pre-trained self-supervised speech model as the front end. They studied the combination of different back-end architectures and self-supervised front ends, as well as the performance of self-supervised models using different pre-training methods, and proved that fine-tuning could achieve better results. Tak et al. [18] applied the wav2vec2.0 front-end, fine-tuning form of self-supervised learning and data enhancement to audio deepfake detection, which improved the generalization and robustness with better results.

Detection performance can be significantly improved using a DNN front end based on traditional features and training on a standard database. However, when facing real and faked speech in the unknown domain, some information will be lost. Furthermore, it is usually only effective for detecting specific types of faked speech, and the detection performance will be reduced. Training a robust and generalized DNN-based front-end feature extractor requires a large amount of natural and fake speech data. Moreover, these DNN-based front ends are trained in a supervised method, which requires many human and material resources. Therefore, this paper proposes an audio deepfake detection method, HuRawNet2_modified, based on a self-supervised pre-training model.

3. Methods

The flow chart of the HuRawNet2_modified method is shown in Figure 1. The whole model consists of a pre-trained HuBERT-based and back-end detection model. The input of the entire model is the original waveform, and the output is the result of binary classification. Firstly, the data were pre-processed by adding the impulse signal and white noise additive noise to the original audio for data enhancement (see Section 3.1 for details). Next, a self-supervised pre-trained model and fine-tuning (see Section 3.2 for more information) were used to extract acoustic features. A fully connected layer was added after the self-supervised front end to train jointly with the back-end detection model and reduce the dimensionality of the self-supervised model output. The extracted acoustic features were then processed by the three residual blocks of the back-end detection model (see Section 3.3 for details), where α-FMS was used to obtain more discriminative features. Finally, a softmax activation function was used in the output layer to obtain real or fake detection results.

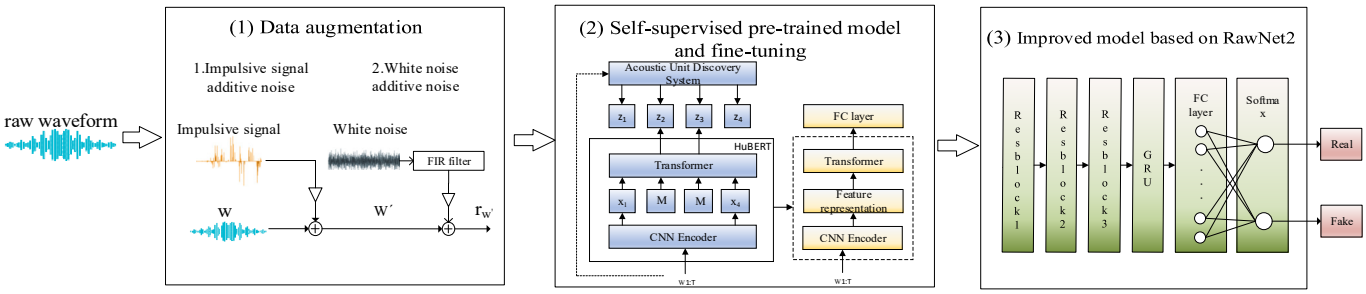


Figure 1. HuRawNet2_modified audio deepfake detection method.

3.1. Data Augmentation

Data augmentation (DA) is often used in machine learning tasks to generate new data from the dataset. The added data were used for training, which can help reduce overfitting and bias, thereby improving classification performance. Some data enhancement methods have been proposed and applied to audio deepfake detection, and SpecAugment is widely used [19]. SpecAugment is a spectrum augmentation method, which is only suitable for the audio deepfake model based on front-end feature extraction, and it is not easy to operate on the audio waveform. This study used two methods of impulsive signal additive noise and white noise additive noise to enhance the data in series. These methods do not require additional data or modifications to the model and they operate directly on the original waveform, which can be appropriate for downstream tasks.

Impulsive signal noise, also known as salt and pepper noise, is discontinuous and consists of irregular pulses or noise spikes with short duration and a large amplitude [20]. The disturbance was applied to the sample to obtain Equation (1):

$$w'[i] = w[i] + z_w[i], \tag{1}$$

where w represents the original audio with L samples.

The use of signal-independent additive noise is one of the common forms of data augmentation, which has been applied to various tasks, such as speech recognition, speaker recognition, etc. The power of white noise in each frequency band was evenly distributed, processed by the FIR filter, and added to the speech, as shown on the right side of data augmentation in Figure 1. The equation is shown in Equation (2):

$$r_{w'}[i] = w'[i] + \frac{10^{\frac{SNR}{20}}}{\|z_{w'}\|^2 \cdot \|w'\|^2} \cdot z_{w'}[i], \tag{2}$$

where w' represents the audio added with impulsive additive noise, SNR refers to the random signal-to-noise ratio, $SNR \in [10, 40]$, $z_{w'}$ denotes the result of white noise after FIR filter processing, and $r_{w'}$ denotes the result after data pre-processing.

3.2. Self-Supervised Pre-Training Speech Models and Fine-Tuning

The distinguishing features in the front end of traditional detection systems usually use well-designed hand-crafted manual features. The detection performance of a detection system fundamentally depends on the extracted features. Nevertheless, traditional hand-crafted features lose some information and are usually effective only for detecting specific types of deepfake speech, affecting the system's generalizability. There can be many labels for speech, such as speakers, words, phonemes, etc. If only one of the labels is used for learning, the learned model performance is insufficient. However, self-supervised learning can be unaffected, which gives the self-supervised learning model excellent generalizability. When only a small amount of labeled data are used to learn the classifier from the result representation, the pre-training model using self-supervised learning can be applied effectively to many different tasks [10].

3.2.1. Self-Supervised Pre-Trained Speech Model

WenetSpeech is currently the largest open-source Mandarin speech corpus with transcriptions. The data were mainly derived from YouTube videos and Podcast audio, covering various types of recording scenes, background noise, speaking methods, etc. It specifically included 10 scenes such as audiobooks, commentary, documentaries, TV shows, interviews, etc., with more than 10,000 h of data [16]. Among the many self-supervised speech models, this study used the HuBERT pre-training model trained on the WenetSpeech Chinese dataset, and the HuBERT pre-training model trained on the LibriSpeech English dataset, to test the effect on the FMFCC-A Chinese dataset [8] and the ASVspoof 2021 LA English dataset.

Let W denotes a speech utterance $W = [w_1, w_2, \cdots, w_T]$ of T-frames. As is shown in Figure 2a, the acoustic unit discovery system generates the target label $f(x) = Z = \{z_1, z_2, \cdots, z_T\}$ with the k-means clustering algorithm, such as MFCC features. At the same time, W generates a feature sequence $[x_1, x_2, \cdots, x_T]$ through a CNN encoder. Let M be the index of the masked sequence X', and X' represents the masked sequence, using the same strategy as wav2vec 2.0 to generate the mask, $X' = random(X, M)$. The variable p represents the proportion of randomly selected starting indices throughout the entire T-frame of speech. The variable l represents the step size, which is set to 10 and $M = p\% \times T + l$. If $t \in M$; then x_t is replaced by an embedded mask, and the mask prediction model G takes X' as the input and predicts the label distribution of the discrete units.

The BERT encoder is composed of many layers of transformer encoders. This study used a BERT encoder consisting of 12 layers of transformer encoders. The BERT encoder inputs the mask sequence X' and outputs a feature sequence $O = [o_1, o_2, \cdots, o_T]$. The proportion p of the start index in the entire T-frame of speech is as per that shown in Equation (3):

$$p_g^{(k)}(c \mid \widetilde{X}, t) = \frac{\exp\left(\text{siminarity}\left(\mathbf{A}^{(k)} o_t, \boldsymbol{e}_c\right)/\tau\right)}{\sum_{c'=1}^{C} \exp\left(\text{siminarity}\left(\mathbf{A}^{(k)} o_t, \boldsymbol{e}_{c'}\right)/\tau\right)}, \tag{3}$$

where $\mathbf{A}$ is the projection matrix, e_c is the embedding for the codebook, similarity calculates the cosine similarity between two vectors, and τ is used to scale the logarithmic function, which is set to 0.1.

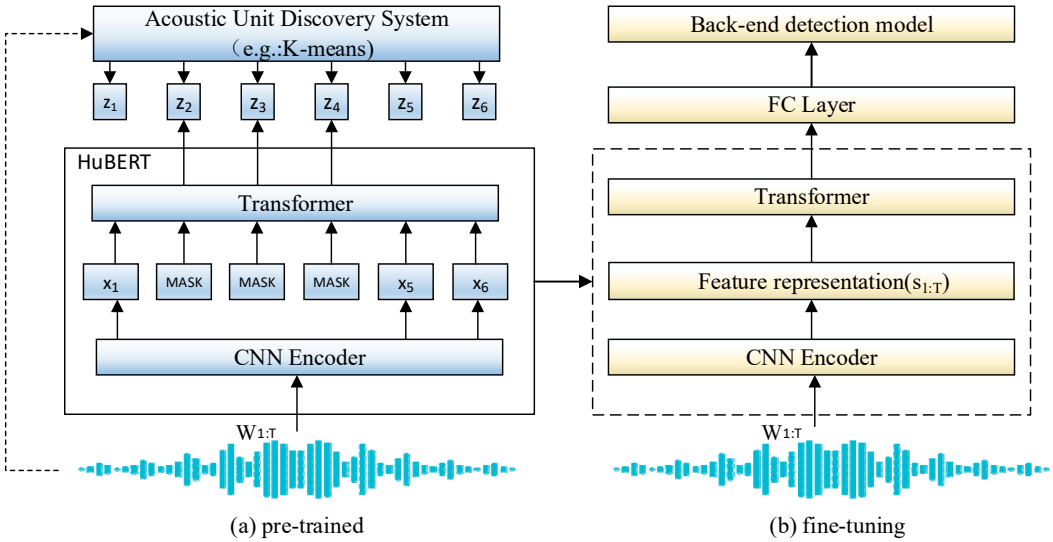


Figure 2. The overall structure of pre-training and fine-tuning.

By iterative training, the cross-entropy loss functions $LOSS_{mask}$ and $LOSS_{unmask}$ are calculated on masked and unmasked units, and then the final loss value LOSS is obtained by weighted summation, as shown in Equation (4):

$$LOSS = \alpha LOSS_{Mask} + (1 - \alpha) LOSS_{unmask}, \quad (4)$$

The HuBERT model is similar to the classic wav2vec 2.0 model, but the training methods are different. The latter is to discretize the audio features as a self-supervised target during training, which is characterized by calculating the loss function only in the mask area; the former obtains the training target by carrying out k-means clustering on MFCC features. The results show that the performance of the HuBERT model is better than that of the wav2vec 2.0 model [9].

3.2.2. Fine-Tuning

Fine-tuning is one of the transfer learning methods suitable for smaller datasets and it has low training costs, which can improve the detection performance for known attacks. Some studies have shown that fine-tuning is beneficial and can prevent overfitting, promoting better generalization [14]. Pre-training only extracts features of natural speech, and fine-tuning, with both natural and deepfake audio data, enables the self-supervised pre-training model to adapt to the downstream task of audio deepfake detection, which helps to improve detection performance.

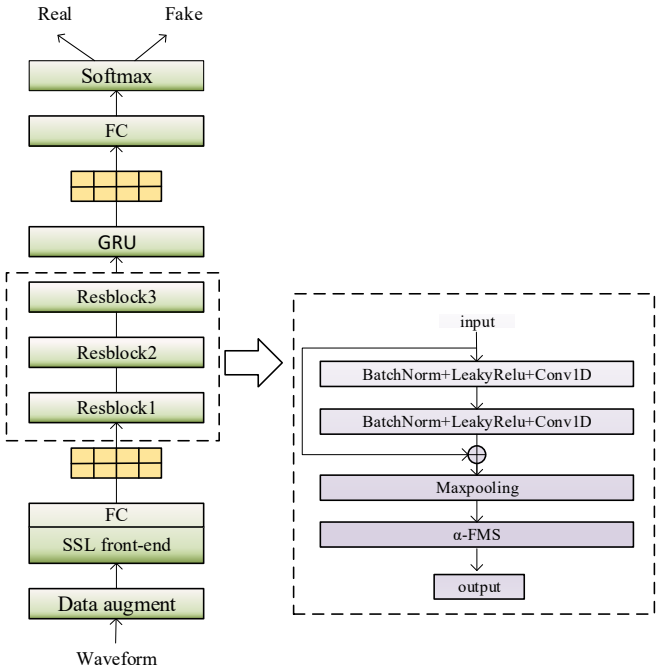
The process of fine-tuning is shown in Figure 2b. After pre-training on unlabeled data, fine-tuning was performed on the two training sets with labels. The back-end detection model and the pre-trained HuBERT model were jointly optimized by back-propagation, and the weighted cross-entropy loss function was used to calculate the loss. The speech $W = [w_1, w_2, \cdots, w_T]$ of the T-frame was passed through the CNN encoder to obtain the potential speech representation $S = [s_1, s_2, \cdots, s_T]$. and then sent to the transformer encoder to obtain the context representation R. In order to reduce the dimension, a fully connected layer (FC layer) was added after the output of the transformer encoder, and the output was sent to the residual blocks of the back-end detection model. In this paper, the ASVspoof 2021 LA training set (same as ASVspoof 2019 LA training set) and the FMFCC-A training set were used for fine-tuning, and the detection performance of the model was tested on different evaluation sets.

3.3. Improved Model Based on RawNet2

With the development of deep learning, models that operate directly on the raw waveform are becoming more common. Most existing work uses a convolutional layer or sinc filter to process the raw waveform input. RawNet2 was a novel end-to-end network model proposed by Jung et al. [21] in 2020, and applied to audio deepfake detection by Tak et al. [18] in 2021, with good results. It has been set as the baseline system for the ASVspoof 2021 challenge.

The back-end detection model in this study was based on RawNet2 and consisted of residual blocks, a gate recurrent unit, a fully connected layer, and an output layer. The input feature sequence was first extracted from the frame-level representation by residual blocks. Then, the gate recurrent unit (GRU) was used to aggregate the frame-level representation into an utterance-level representation for the analysis and discrimination of the entire sequence, which was then fed into the fully connected layer. When the trained model was used in the evaluation set, a softmax activation function was added after the fully connected layer to obtain real or deepfake detection results. The real speech label was 1 and the deepfake speech label was 0. The classification effect was evaluated with a threshold of 0.5.

The original Rawnet2 model cannot fully extract the deeper features of fake audio, cannot effectively distinguish the key features of real and deepfake speech, and the generalizability of the model needs to be improved. Therefore, this study made the following improvements to the RawNet2 model: (1) A self-supervised speech pre-training model was used instead of sinc convolutional layers; (2) It had an improved residual structure with α-FMS instead of FMS; (3) The number of residual blocks were reduced. Most end-to-end speaker recognition models have degraded performance compared to models using manual features, while the widely adopted ECAPA-TDNN model and its variants [22,23] enable an EER below 1%. In this study, we followed the setting of the ECAPA-TDNN model and reduced the number of residual blocks from 6 to 3 to speed up the training and make the model more efficient. The structure of the improved model is shown in Figure 3a, and the structure of the improved residual block is shown in Figure 3b.



(a) structure of RawNet2_modified (b) structure of residual block

Figure 3. Improved model framework based on RawNet2.

FMS [21] (filter-wise feature map scaling), used in residual blocks, refers to filter-based feature map scaling. The purpose of FMS is to modify the size of a given feature map independently, the output of the residual block, to obtain a more discriminative representation and improve the performance and generalization of the model, with the advantages of reducing model parameters and computation. FMS obtains scaling vectors from feature mapping and then adds or multiplies them with features or applies these two operations in turn, as shown in reference [21]. The multiplicative FMS is similar to the attention map for the attention mechanism, but uses the sigmoid activation function instead of the softmax function. This is because using the softmax function may cause the information to be over-removed. However, the limitations are that FMS uses the same scaling vector for addition and multiplication, can only add values between 0 and 1 during addition, and has difficulty optimizing addition and multiplication operations simultaneously when performing multiplication.

To solve this problem, Jung et al. [24] improved it and proposed α-FMS. A trainable parameter α is added to each filter and multiplied by the scaling vector. The parameter α is automatically learned by back propagation and optimization algorithms during training. Each filter has its scaling vector, which can further improve the performance and generalizability of the model compared to FMS. The specific operation is shown in Equation (5).

As is shown in the α-FMS structure diagram in Figure 4, let $C = [C_1, C_2, \cdots, C_F]$ be the feature map of the residual block, $C_f \in \mathbb{R}^T$, T be the length of the time series, and F be the number of filters. The scaling vector is first obtained by performing global average pooling on the time axis, then feedforward through a fully connected layer, and finally sigmoid activation. Let $S = [S_1, S_2 \cdots, S_F]$ be the scaling vector, $C' = [C'_1, C'_2, \cdots, C'_F]$ be the scaled feature map, $S_f \in \mathbb{R}^1$, $C'_f \in \mathbb{R}^T$, S_f and C_f are copied to perform element-by-element operations. The purpose of additive FMS is to add a slight disturbance to the feature map to increase the discriminative power of the feature map [25]. Add α for each filter in Equation (5).

$$C'_f = (C_f + \alpha) \times S_f \tag{5}$$

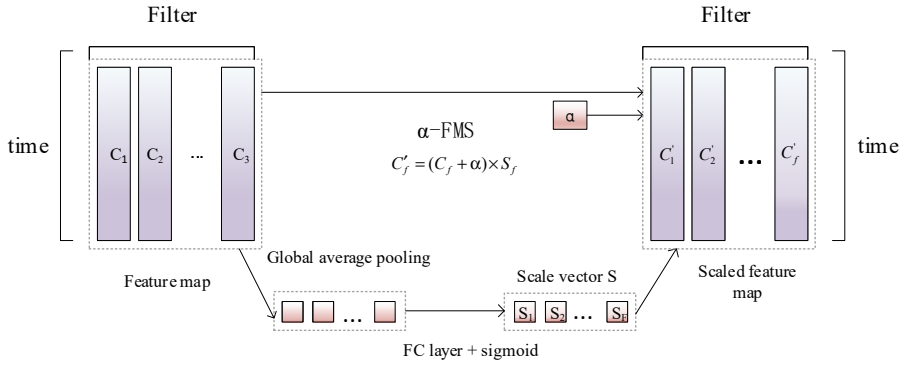


Figure 4. Structure of α-FMS.

One advantage of this method is that it allows the model to autonomously learn the most suitable feature map scaling ratio for the task, rather than being manually set. This can enhance the expressive power and flexibility of the model, thereby improving the model's performance.

4. Experiment

This section describes the dataset used, the evaluation metrics, and the experimental results and analysis to train the binary classifier for the ASVspoof 2021 LA dataset and the FMFCC-A dataset, respectively, for distinguishing the results as natural or faked speech. Fine-tuning requires a large amount of GPU memory, so the voice data were processed into approximately four seconds of speech; those that were longer than four seconds were

cut, and those that were less than four seconds were first copied for the speech before processing. In this study, the experimental iteration number Epoch was 40, using the Adam optimizer and default settings. When the sinc filter was used, the learning rate was fixed at 0.0001; when a self-supervised front end was used, the fine-tuning demanded high computer computation, and the learning rate was chosen to be initialized at 0.000001 and adjusted by the cosine annealing learning rate decay with the batch size of 14 to avoid overfitting due to the experimental conditions. The experimental environment of this study is shown in Table 1. A DCU (deep computing unit) is an accelerator card dedicated to AI (artificial intelligence) and deep learning.

Table 1. Experimental environment.

Name	Version
CPU	C86 7185 32-core Processor 2.0 GHz
Accelerator card	Dcu2
Memory	16 GB
Operating system	CentOS Linux 7.6 64-bit
Python	3.7.2
Deep learning library	PyTorch 1.10.0, fairseq 0.10.0

4.1. Datasets and Evaluation Metrics

This study included experiments on the English ASVspoof 2021 LA dataset, Asvspoof 2019 LA dataset, Chinese FMFCC-A dataset, and Chinese FAD dataset. All four datasets were divided into three parts: training set, development set, and evaluation set, and the speakers in the subsets of the same dataset did not overlap with each other. The ASVspoof 2019 LA dataset is from 107 different speakers and contains real and fake discourse generated using 17 different TTS and VC systems [6]. The training and development sets for the ASVspoof 2021 LA dataset are the same as those released for the ASVspoof 2019 challenge. The evaluation set was recorded by 48 speakers corresponding to the ASVspoof 2019 challenge evaluation set [7].

The FMFCC-A dataset contains a collection of 40,000 synthetic and 10,000 genuine utterances. Moreover, the fake audios was generated based on 11 Mandarin TTS systems and 2 Mandarin VC systems, and the duration is randomly set in the range between 2 and 10 s, with the sampling rate of 16 kHz [8]. The FAD dataset consists of 12 types mainstream voice deepfake techniques such as STRAIGHT, LPCNet, and HifiGAN to generate fake audios, and real audios from six different corpora such as AISHELL1, AISHELL3, and THCHS-30 [26]. The evaluation set of the FAD dataset contains 14,000 utterances generated by four unknown deepfake methods that were not included in the training and validation sets, which can better detect the robustness and generalization of the model in the face of unknown attacks. The specific information of the dataset used in this paper is shown in Table 2.

Table 2. Details of three databases.

Database	Number of Utterances							Language	Storage Format
	Total	Train		Development		Evaluation			
		Real	Fake	Real	Fake	Real	Fake		
ASVspoof 2021 LA	231,790	2580	22,800	2548	22,296	14,816	166,750	English	FLAC
ASVspoof 2019 LA	121,461	2580	22,800	2548	22,296	7355	63,882	English	FLAC
FMFCC-A	50,000	4000	6000	3000	17,000	3000	17,000	Chinese	WAV
FAD	115,800	12,800	25,600	4800	9600	21,000	42,000	Chinese	WAV

To evaluate the performance of the detection system, this study used two evaluation metrics commonly used for audio deepfake detection: equal error rate (EER) and tandem detection cost function (t-DCF) as evaluation metrics. The min t-DCF was proposed by

the ASVspoof 2019 challenge and improved by the ASVspoof 2021 challenge. The specific formulae for EER and min $t-DCF$ are as follows:

$$P_{\text{false}}(\theta) = \frac{faked\ voice\ with\ score > \theta}{total\ faked\ voice}, \tag{6}$$

$$P_{\text{miss}}\ (\theta) = \frac{true\ voice\ with\ score \leq \theta}{total\ true\ voice} \tag{7}$$

$$EER = P_{\text{false}}(\theta_{EER}) = P_{\text{miss}}(\theta_{EER}) \tag{8}$$

$$\min\ t-DCF = \min_{\theta}\left\{C_0 + C_1 P_{miss}(\theta) + C_2 P_{false}(\theta)\right\} \tag{9}$$

where the EER denotes the error rate when the false alarm rate $P_{false}(\theta)$ and the miss alarm rate $P_{miss}(\theta)$ are equal, and θ_{EER} denotes the threshold value when $P_{false}(\theta)$ and $P_{miss}(\theta)$ are equal. The smaller the EER, the better the performance of the detection system. The smaller the t-DCF, the better the generalizability of the detection system and the smaller the impact of the performance of an ASV system [27].

The $Log-loss$ function ($Log-loss$) is also used as an evaluation metric for the FMFCC-A dataset. $Log-loss$ is one of the primary metrics used to evaluate the performance of a classification problem, indicating how close the predicted probability is to the corresponding actual value.

$$Log-loss_i = -[y_i \ln p_i + (1-y_i)\ln(1-p_i)] \tag{10}$$

where i denotes the index of the statement, y_i is the corresponding label, and p_i is the predicted probability. When the $Log-loss$ is smaller, the predicted probability is closer to the true value, and the model performance is better.

4.2. Experimental Results and Analysis

In this study, we used the HuBERT pre-trained models trained on the WenetSpeech Chinese dataset and LibriSpeech English dataset, and fine-tuned them using ASVspoof 2021 LA training set, FMFCC-A training set, and FAD training set, to examine the performance of the ASVspoof 2021 LA evaluation set, ASVspoof 2019 LA evaluation set, FMFCC-A development set, and FAD evaluation set.

4.2.1. Comparison Experiments

The EER and min t-DCF of the baseline model and HuRawNet2_modified method for the evaluation set of the ASVspoof 2021 challenge on the Asvspoof 2021 dataset are shown in Table 3. "WenetSpeech" indicates pre-training on WenetSpeech Chinese dataset, and "LibriSpeech" indicates pre-training on LibriSpeech English dataset.

Table 3. Performance for the ASVspoof 2021 evaluation partition in terms of EER (%) and min t-DCF.

Model	EER (%)	Min t-DCF
CQCC-GMM [28]	15.62	0.4974
LFCC-GMM [29]	19.30	0.5758
LFCC-LCNN [30]	9.26	0.3445
Raw audio-RawNet2 [3]	9.50	0.4257
HuRawNet2_modified (LibriSpeech)	2.89	0.2182

The analysis of Table 3 shows that the detection performance of the proposed method was significantly improved compared with the other four baseline models in the ASVspoof 2021 competition. Compared with Baseline RawNet2, the EER and min t-DCF indicators were reduced by 69.5% and 48.7%, respectively. This proves that the method of using

the self-supervised pre-training model to extract general features and then fine-tuning is more suitable for audio deepfake detection tasks. It can be seen that the self-supervised pre-training model was of practical value. The EER and $Log-loss$ of the baseline models and HuRawNet2_modified method for the development set of the FMFCC-A dataset of the second Fake Media Forensic Challenge of CSIG are shown in Table 4.

Table 4. Performance for the FMFCC-A evaluation partition in terms of EER (%) and $Log-loss$.

Model	EER (%)	Log-Loss
CQCC-ResNet34 [31]	7.27	0.5398
Raw audio-Res-TssDNet [2]	8.26	0.8152
HuRawNet2_modified (WenetSpeech)	3.25	0.3121

To verify the performance of this model on Chinese deepfake speech, the FMFCC-A and FAD datasets have been introduced in this paper. Analysis of the results in Table 4 shows that the EER of the model proposed in this paper was reduced by 55.3% and 60.7%, and the $Log-loss$ was reduced by 42.2% and 61.7% for the FMFCC-A dataset and the FAD dataset, respectively, compared to the two baseline systems. This indicated that the performance of the model had been improved and the detection performance of Chinese faked speech was better. The LCNN-LSTM model, among the four baseline models of the ASVspoof2021 challenge, achieved the best result on the FAD dataset (EER = 13.91), and the other specific results are not displayed in this paper. As compared to the LCNN-LSTM model, the EER of HuRawNet2_modified was reduced by approximately 29.9% and 40.25% for the FMFCC-A dataset and the FAD dataset, respectively. Therefore, we can conclude that the performance was improved compared to the baseline model on different datasets, indicating that HuRawNet2_modified model has a better generalizability and a certain advantage in terms of detection performance.

In order to verify the above conclusions, this study used four datasets for testing, and the results are shown in Table 5 and Figure 5. The results of pre-training using different language datasets on the three datasets showed that the EER and min t-DCF were slightly reduced when the pre-trained dataset, and the fine-tuned and tested dataset were in the same language.

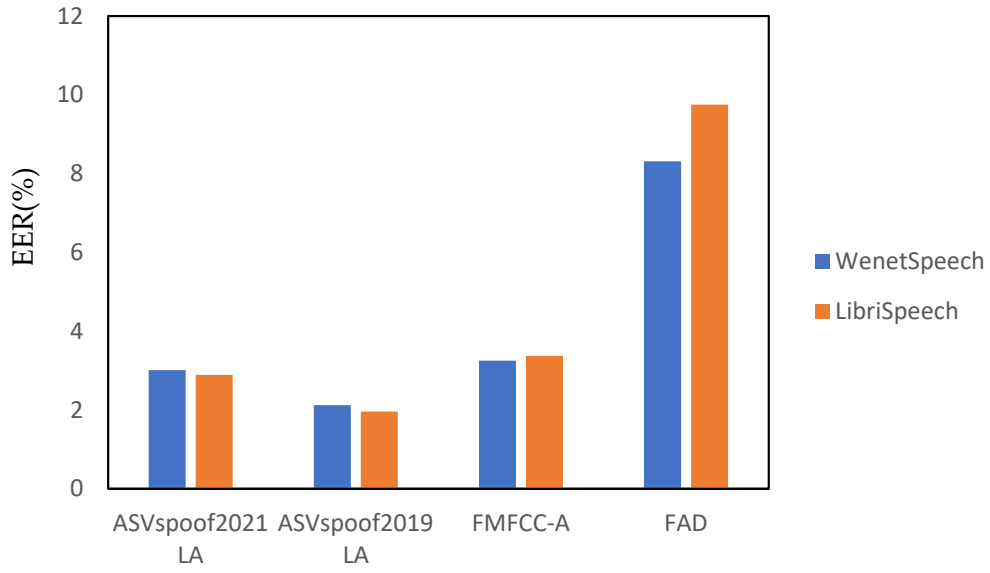


Figure 5. Results of HuRawNet2_modified model on different datasets using different language pre-trained models.

Table 5. Results of HuRawNet2_modified model on different datasets using different language pre-trained models.

Dataset	EER (%)	Min t-DCF
ASVspoof 2021 LA (WenetSpeech)	3.01	0.2278
ASVspoof 2021 LA (LibriSpeech)	2.89	0.2182
ASVspoof 2019 LA (WenetSpeech)	2.12	0.1442
ASVspoof 2019 LA (LibriSpeech)	1.96	0.1393
FMFCC-A (WenetSpeech)	3.25	0.3121
FMFCC-A (LibriSpeech)	3.37	0.3378
FAD (WenetSpeech)	8.31	0.1730
FAD (LibriSpeech)	9.75	0.2292

A self-supervised speech model pre-trained on different language datasets, using the same network model, was used to extract features, and fine-tune and test different language datasets with different detection performances. As is shown in rows 1 and 2 of Table 5, fine-tuning and detection on the ASVspoof 2021 dataset provided a gain of approximately 4.1% to the model, when using the pre-trained model trained on the LibriSpeech English dataset over the WenetSpeech Chinese dataset. As is shown in rows 3 and 4 of Table 5, fine-tuning and testing on the ASVspoof 2019 dataset provided a gain of approximately 7.5% to the model when using the pre-trained model trained on the LibriSpeech English dataset than when using the WenetSpeech Chinese dataset. As is shown in rows 5 and 6 of Table 5, fine-tuning and detection on the FMFCC-A dataset provided a gain of approximately 3.6% to the model when using the pre-trained model trained on the WenetSpeech Chinese dataset over the LibriSpeech English dataset. When experimenting on the FAD dataset, the pre-trained model using the same language could be improved by approximately 14.7%. The following conclusions can be visualized more clearly in Figure 5. Based on the experimental results, it can be tentatively demonstrated that acoustic features can be extracted across languages using a self-supervised pre-trained speech model with fine-tuning. However, the detection effect can be slightly improved when the pre-trained, and the fine-tuned and tested datasets are in the same language.

4.2.2. Ablation Experiments

To verify the improvement of detection performance, this study adopted the ablation experiments on the ASVspoof 2021 LA dataset, using the RawNet2 model as the base network (row 4 in Table 3). Data augmentation, the α-FMS module, and a self-supervised speech pre-training and fine-tuning module were gradually added, and they were compared with the sinc filter of RawNet2. The experimental results are shown in Table 6 ("--" means the method is not included, "$\surd$" means the method is included). Figure 6 plots the performance results of the ablation experiments. The closer the data are to the origin of the coordinates, the better the detection effect and generalizability.

Table 6. Ablation experiments of HuRawNet2_modified.

Abbreviation	Method				Metric	
	Front End		Data Augmentation	α-FMS	EER (%)	Min t-DCF
	SSL Pre-Trained and Fine-Tuned	Sinc Filter				
SSL	$\surd$	--	--	--	5.49	0.3687
+DA	$\surd$	--	$\surd$	--	4.89	0.3357
+α-FMS	$\surd$	--	--	$\surd$	4.51	0.3268
+DA+α-FMS	$\surd$	--	$\surd$	$\surd$	2.89	0.2182
Sinc	--	$\surd$	--	--	10.17	0.5103
+DA	--	$\surd$	$\surd$	--	9.19	0.4761
+α-FMS	--	$\surd$	--	$\surd$	8.25	0.4573
+DA+α-FMS	--	$\surd$	$\surd$	$\surd$	5.52	0.3964

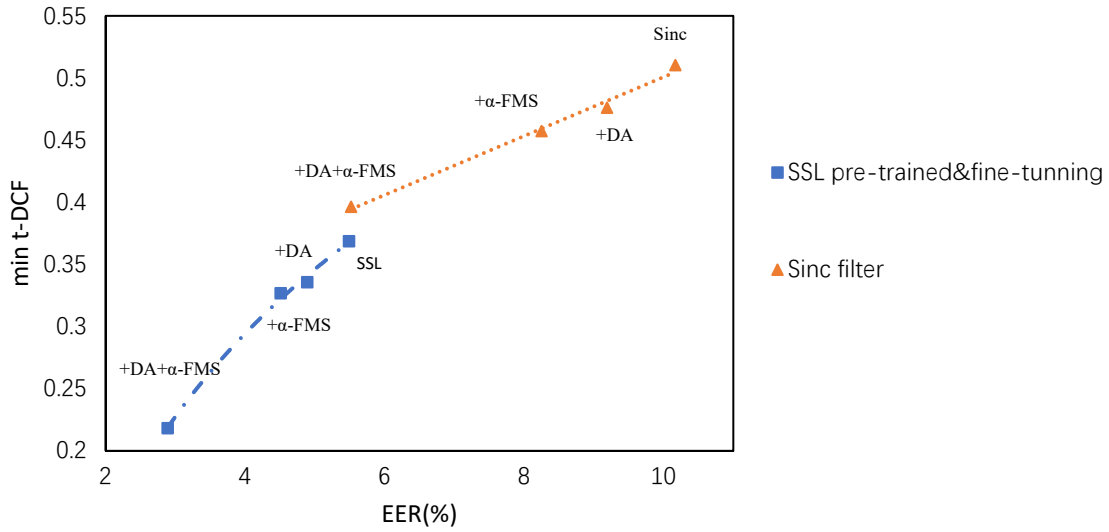


Figure 6. The results of ablation experiments.

In terms of EER for comparison, it can be seen from rows 2 and 6 of Table 6 that adding the data augmentation module resulted in a 10.6~12.3% gain to the model; adding only α-FMS in rows 3 and 7, the model performance was improved by approximately 17.8~18.9%; adding both data augmentation and α-FMS, the EER was reduced by approximately 45.7~47.3% from rows 4 and 8, indicating that both data augmentation and α-FMS contribute to the model performance improvement, with α-FMS adding more to the model. From Figure 6, it can be seen that the self-supervised pre-training and fine-tuning front end was closer to the origin than the sinc filter's front end. It can be concluded that the EER and min t-DCF of the method based on self-supervised pre-training and fine-tuning proposed in this paper keep decreasing. The results were generally better than those of the front end of the sinc filter, which further verifies that the method proposed in this paper can fully extract deepfake speech features, and improve the detection effect and generalizability compared with mainstream detection algorithms.

5. Conclusions and Discussion

In this study, we designed an audio deepfake detection model based on a self-supervised pre-training model with improvements in two parts: front-end feature extraction and a back-end classification model. For front-end feature extraction, the model performed data augmentation using a self-supervised model to extract generic linguistic features, which were then fine-tuned in two separate datasets. On the back-end classification model, RawNet2 was improved by introducing α-FMS to enhance the discrimination of the feature maps. With the continuous development of deepfake technology, detection technology will face a more severe test. In future studies, subsequent attempts will be made to integrate the system with the speech recognition system, improve generalizability, study the gains associated with different loss functions, and focus on the effectiveness of detecting various attacks.

Subsequent research will concentrate on unlabeled detection techniques, encompassing methods such as self-supervised knowledge distillation, aiming to yield more pragmatically applicable solutions. We believe that this delineates one of the future trajectories for deep learning research. Confronted with the escalation of innovative and unfamiliar forgery methodologies, our objective remains steadfast in preserving the ability to detect audio forgeries, thus augmenting the model's generalization capability. Furthermore, the forthcoming ASVspoof 2023 challenge is on the horizon, prompting us to continually monitor and appreciate the advancements achieved in the realm of voice deepfake detection.

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

Funding: This research was supported by the Double First-Class Innovation Research Project for People's Public Security University of China (No.2023SYL07).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The ASVspoof 2021 dataset can be found at https://www.asvspoof.org/index2021.html (accessed on 28 May 2023). The ASVspoof 2019 dataset can be found at https://www.asvspoof.org/index2019.html (accessed on 28 May 2023). The FMFCC-A dataset can be found at https://github.com/Amforever/FMFCC-A (accessed on 28 May 2023). The FAD dataset can be found at https://zenodo.org/record/6635521#.Ysjq4nZBw2x (accessed on 28 May 2023).

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

References

1. Wang, X.; Yamagishi, J. A Practical Guide to Logical Access Voice Presentation Attack Detection. In *Frontiers in Fake Media Generation and Detection*; Springer Nature: Singapore, 2022; pp. 169–214. [CrossRef]
2. Hua, G.; Teoh, A.B.J.; Zhang, H. Towards End-to-End Synthetic Speech Detection. *IEEE Signal Process. Lett.* **2021**, *28*, 1265–1269. [CrossRef]
3. 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.
4. Tak, H.; Jung, J.; 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.
5. Wu, Z.; Kinnunen, T.; Evans, N.; Yamagishi, J.; Hanilçi, C.; Sahidullah, M.; Sizov, A. ASVspoof 2015: The First Automatic Speaker Verification Spoofing and Countermeasures Challenge. In Proceedings of the Interspeech 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 4–5 September 2015; ISCA: Dresden, Germany, 2015; pp. 2037–2041.
6. Nautsch, A.; Wang, X.; Evans, N.; Kinnunen, T.H.; Vestman, V.; Todisco, M.; Delgado, H.; Sahidullah, M.; Yamagishi, J.; Lee, K.A. ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech. *IEEE Trans. Biom. Behav. Identity Sci.* **2021**, *3*, 252–265. [CrossRef]
7. Yamagishi, J.; Wang, X.; Todisco, M.; Sahidullah, M.; Patino, J.; Nautsch, A.; Liu, X.; Lee, K.A.; Kinnunen, T.; Evans, N.; et al. ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection. In Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, Online, 16 September 2021; ISCA: Dresden, Germany, 2021; pp. 47–54.
8. Zhang, Z.; Gu, Y.; Yi, X.; Zhao, X. FMFCC-a: A Challenging Mandarin Dataset for Synthetic Speech Detection. In Proceedings of the Digital Forensics and Watermarking: 20th International Workshop, Beijing, China, 20–22 November 2021; Springer: Beijing, China, 2022; pp. 117–131.
9. Hsu, W.-N.; Bolte, B.; Tsai, Y.-H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEEACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [CrossRef]
10. Kenton, J.D.M.-W.C.; Toutanova, L.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, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
11. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. In Proceedings of the Advances in Neural Information Processing Systems 26 (NIPS 2013), Lake Tahoe, NV, USA, 5–10 December 2013.
12. Van den Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv* **2018**, arXiv:1807.03748.
13. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. Wav2vec: Unsupervised Pre-Training for Speech Recognition. In Proceedings of the Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15–19 September 2019; pp. 3465–3469.
14. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Proceedings of the Advances in Neural Information Processing Systems 2020, Virtual, 6–12 December 2020; Curran Associates, Inc.: New York, NY, USA, 2020; Volume 33, pp. 12449–12460.

15. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. [CrossRef]
16. Zhang, B.; Lv, H.; Guo, P.; Shao, Q.; Yang, C.; Xie, L.; Xu, X.; Bu, H.; Chen, X.; Zeng, C.; et al. WENETSPEECH: A 10,000+ Hours Multi-Domain Mandarin Corpus for Speech Recognition. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 7–13 May 2022; pp. 6182–6186.
17. Wang, X.; Yamagishi, J. Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures. In Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2022), Beijing, China, 28 June–1 July 2022; ISCA: Beijing, China, 2022.
18. Tak, H.; Todisco, M.; Wang, X.; Jung, J.; Yamagishi, J.; Evans, N. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation. In Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2022), Beijing, China, 28 June–1 July 2022.
19. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In Proceedings of the Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15–19 September 2019; pp. 2613–2617.
20. Zhang, J.; Qiu, T.; Luan, S. An Efficient Real-Valued Sparse Bayesian Learning for Non-Circular Signal's DOA Estimation in the Presence of Impulsive Noise. *Digit. Signal Process.* **2020**, *106*, 102838. [CrossRef]
21. Jung, J.; Kim, S.; Shim, H.; Kim, J.; Yu, H.-J. Improved Rawnet with Feature Map Scaling for Text-Independent Speaker Verification Using Raw Waveforms. In Proceedings of the Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Shanghai, China, 25–29 October 2020; pp. 1496–1500.
22. Desplanques, B.; Thienpondt, J.; Demuynck, K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In Proceedings of the Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Shanghai, China, 25–29 October 2020; pp. 3830–3834.
23. Ravanelli, M.; Parcollet, T.; Plantinga, P.; Rouhe, A.; Cornell, S.; Lugosch, L.; Subakan, C.; Dawalatabad, N.; Heba, A.; Zhong, J.; et al. SpeechBrain: A General-Purpose Speech Toolkit. *arXiv* **2021**, arXiv:2106.04624.
24. Jung, J.; Shim, H.; Kim, J.; Yu, H.-J. α-feature map scaling for raw waveform speaker verification. *J. Acoust. Soc. Korea* **2020**, *39*, 441–446.
25. Zhang, J.; Inoue, N.; Shinoda, K. I-Vector Transformation Using Conditional Generative Adversarial Networks for Short Utterance Speaker Verification. In Proceedings of the Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018; ISCA: Hyderabad, India, 2018.
26. Ma, H.; Yi, J.; Wang, C.; Yan, X.; Tao, J.; Wang, T.; Wang, S.; Xu, L.; Fu, R. FAD: A Chinese Dataset for Fake Audio Detection. *arXiv* **2022**, arXiv:2207.12308.
27. Jung, J.; Heo, H.-S.; Tak, H.; Shim, H.; 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), Virtual, 7–13 May 2022; pp. 6367–6371.
28. Todisco, M.; Delgado, H.; Evans, N. Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification. *Comput. Speech Lang.* **2017**, *45*, 516–535. [CrossRef]
29. Sahidullah, M.; Kinnunen, T.; Hanilçi, C. A Comparison of Features for Synthetic Speech Detection. In Proceedings of the Interspeech 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 4–5 September 2015; ISCA: Dresden, Germany, 2015; pp. 2087–2091.
30. Wang, X.; Yamagishi, J. A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection. In Proceedings of the Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August–3 September 2021; ISCA: Brno, Czechia, 2021; pp. 4259–4263.
31. Todisco, M.; Delgado, H.; Evans, N. A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients. In Proceedings of the Speaker and Language Recognition Workshop, Bilbao, Spain, 21–24 June 2016; ISCA: Bilbao, Spain, 2016; pp. 283–290.