information

Article

Identification of Fake Stereo Audio Using SVM and CNN

Tianyun Liu [1], Diqun Yan [1,*], Rangding Wang [1], Nan Yan [2] and Gang Chen [2]

[1] College of Information Science and Engineering, Ningbo University, Ningbo 315211, China; liutianyun1518@163.com (T.L.); wangrangding@nbu.edu.cn (R.W.)
[2] Ningbo Polytechnic, Ningbo 315800, China; yannan2008h@163.com (N.Y.); bastarcg@163.com (G.C.)
* Correspondence: yandiqun@nbu.edu.cn

Abstract: The number of channels is one of the important criteria in regard to digital audio quality. Generally, stereo audio with two channels can provide better perceptual quality than mono audio. To seek illegal commercial benefit, one might convert a mono audio system to stereo with fake quality. Identifying stereo-faking audio is a lesser-investigated audio forensic issue. In this paper, a stereo faking corpus is first presented, which is created using the Haas effect technique. Two identification algorithms for fake stereo audio are proposed. One is based on Mel-frequency cepstral coefficient features and support vector machines. The other is based on a specially designed five-layer convolutional neural network. The experimental results on two datasets with five different cut-off frequencies show that the proposed algorithm can effectively detect stereo-faking audio and has good robustness.

Keywords: stereo faking audio; audio forensics; MFCC; SVM; CNN

Citation: Liu, T.; Yan, D.; Wang, R.; Yan, N.; Chen, G. Identification of Fake Stereo Audio Using SVM and CNN. *Information* **2021**, *12*, 263. https://doi.org/10.3390/info12070263

Academic Editor: Willy Susilo

Received: 7 May 2021
Accepted: 23 June 2021
Published: 28 June 2021

1. Introduction

Audio forensics [1] is an important branch of multimedia security, which can be used to evaluate the authenticity of digital audio. Many audio forensics methods have been proposed for various speech operations in addition to common audio forgeries, such as double compression [2,3], pitch shifting [4–6], device source [7–9], replaying [10] and the detection of the operation type and sequence of digital speech [11]. Fake-quality detection is a very important part of the field of audio forensics, such as in [12], in which the authors recompressed low bit rate audio into high bit rate audio. Generally, a higher bit rate indicates better audio quality. Since increasing the bit rate will cause recompression, the authors in [12] proposed a fake-quality detection algorithm based on double compression tracing. As far as we know, however, there are no studies related to stereo-faking detection, which also belongs to the category of fake quality. Although fake stereo audio detection remains an unstudied field, there are many similar works in the audio field, such as audio forensics, speech recognition and speaker recognition.

In audio forensics, Mascia et al. [13] proposed a forensic algorithm that uses MFCC/LMSC features to detect the recording environment. The algorithm provides support for forensic analysts in verifying the authenticity of audio content. Vijayasenan et al. [14] proposed a forensic algorithm to study the effect of a wireless channel on physical parameter prediction, based on speech data. Speech data from 207 speakers, along with the corresponding speaker's height and weight, were collected. A Bag of Words (BoW) representation, based on the log magnitude spectrum, was used for features. Support vector regression (SVR) predicted the physical parameters of the speaker from the BoW representation. The proposed system was able to achieve a Root Mean Square Error (RMSE) of 6.6 cm for height estimation, and 8.9 kg for weight estimation for clean speech. Hadoltikar et al. [15] proposed a forensic algorithm for recording device identification, aiming to optimize the MFCC parameters in device identification in an audio recording. Zhao and Malik [16] proposed a forensic algorithm for the identification of the acoustic environment

(acoustic reverberation and background noise are usually used to characterize the acoustic environment). Inverse filtering was used to estimate the reverberation component, and particle filtering was used to estimate background noise from the audio recording. A multi-class support vector machine (SVM) classifier was used for acoustic environment identification (AEI). The experimental results show that the proposed system can successfully identify a recording environment for both regular and blind AEI. Jiang et al. [17] proposed a mobile phone identifier, called Weighted SVM with Weighted Majority Voting (WSVM-WMV), for a closed-set mobile phone identification task. By using Mel-frequency cepstral coefficients (MFCCs) and linear-frequency cepstral coefficients (LFCCs) as the feature vectors, the proposed identifier can improve identification accuracy from 92.42% to 97.86% and from 90.44% to 98.33%, respectively, as compared with the traditional SVM identifier, in identifying a set of 21 mobile phones.

In speech recognition, Mitra et al. [18] proposed an automatic speech recognition (ASR) algorithm, which improves the recognition accuracy of the automatic speech recognition system through the modulation of the medium duration speech amplitude (MMeDuSA) function. Athanaselis et al. [19] discussed the improvement of speech recognition in the presence of noise when a parametric method of signal enhancement is used. Subramanian et al. [20] proposed a method to optimize only a location-guided target speech extraction module along with a speech recognition module with ASR error minimization criteria. Fan et al. [21] proposed a gated recurrent fusion (GRF) method with a joint training framework for robust end-to-end ASR. Compared with the traditional method, the proposed method can reduce the relative character error rate (CER) by 10.04%, using only enhanced functions.

In speaker recognition, Toruk et al. [22] proposed a speaker recognition system based on time-delay neural networks (TDNNs). Compared with the traditional GMM speaker recognition system, its equal error rate (EER) value is improved. However, the relative improvement in the TDNN decreases while the test data duration decreases. Huang [23] proposed speaker characterization using four different data augmentation methods and time-delay neural networks and long–short-term memory neural networks (TDNN-LSTMs). The proposed methods outperform the baselines of both the i-vector and the x-vector. Jagiasi et al. [24] proposed a text-independent and language-independent speaker recognition system that was implemented using dense and convolutional neural networks and explored a system that uses MFCCs along with a DNN and CNN as the models for building a speaker recognition system. Desai et al. [25] found that adding watermarking technology to the speaker's audio can effectively protect the authenticity of the speaker's audio while ensuring its quality. Watermarking technology is 100% efficient in identifying an attack and verifying the authenticity of the speaker.

Stereo faking is one of many audio-quality-faking techniques and aims to convert mono audio into stereo. Mono audio is a single channel of sound, perceived as coming from one position. Stereo audio is the reproduction of sound using two or more independent audio channels in a way that creates the impression of sound from various directions, as in natural hearing. Stereo audio has almost completely replaced mono due to its superior audio quality. Until the 1960s to 1970s, and even earlier, due to the limitations of recording equipment, most audio recordings and movie soundtracks were monophonic, and their auditory effects were poor. To improve audio quality, one can create an artificial stereo audio from its mono version. With the development of audio editing tools, the quality of stereo audio can be very close to real stereo audio. This technique, however, is also manipulated by some criminals for illegal gain. Since 2010, there have been many complaints about fake stereo audio quality, most of which are from music websites and apps. For example, people can buy their favorite songs online. Often, these songs are in stereo format. The legitimate interests of consumers will be violated if the purchased songs are fake stereo created from mono audio.

As shown in Figures 1 and 2, it is difficult to distinguish fake stereo audio from authentic stereo with a waveform or spectrogram, when the mono audio, which is the source of the fake stereo audio, is not given. However, when we listen to fake stereo

audio, we can often feel a significant decline in perceived quality. Hence, it is necessary to design a detection algorithm for fake stereo audio. The detection strategy we adopt is to extract the acoustic features of each channel of stereo audio and combine and feed them to the classifier for recognition. In this work, a corpus for fake stereo audio detection is created from two real stereo datasets. Then, we propose two effective algorithms for detecting fake stereo audio: using (1) the combination of Mel-frequency cepstral coefficients (MFCCs) features and support vector machines (SVMs) (2) the combination of a five-layer convolutional neural network (CNN) and a three-layer fully connected, because SVMs and CNNs have been widely used in audio forensics, speech recognition and other acoustic fields. Wu H. et al. [4] proposed the use of a combination of MFCC features and an SVM classifier to detect electronic disguised voices. Reis et al. [26] used an SVM classifier for audio authentication. Lin et al. [27] used a convolutional neural network to detect audio recapture with high accuracy. Giuseppe Ciaburro [28] used convolutional neural networks to identify the sounds of underground garages. The purpose was to detect sounds to identify dangerous situations and activate automatic alarms to draw attention to the surveillance of the area. This method returned high accuracy in identifying car accidents in underground parking lots. Considering the information of audio sources and channels, eight various classification models are trained. The main contributions of this paper are as follows:

- To our knowledge, this is the first forensic work to identify fake stereo audio in the field of fake quality forensics;
- We provide a corpus for fake stereo audio detection. We have collected a large number of samples from the two most widely used fields (music and film recording) as datasets;
- After studying the method of stereo forgery, we found that the cut-off frequency of different high-pass filters has little effect on the sound quality but can greatly affect the detection effect of the model. The detection algorithm we put forward based on this point has strong robustness, and can detect fake stereo audios with different cut-off frequencies of high-pass filters.

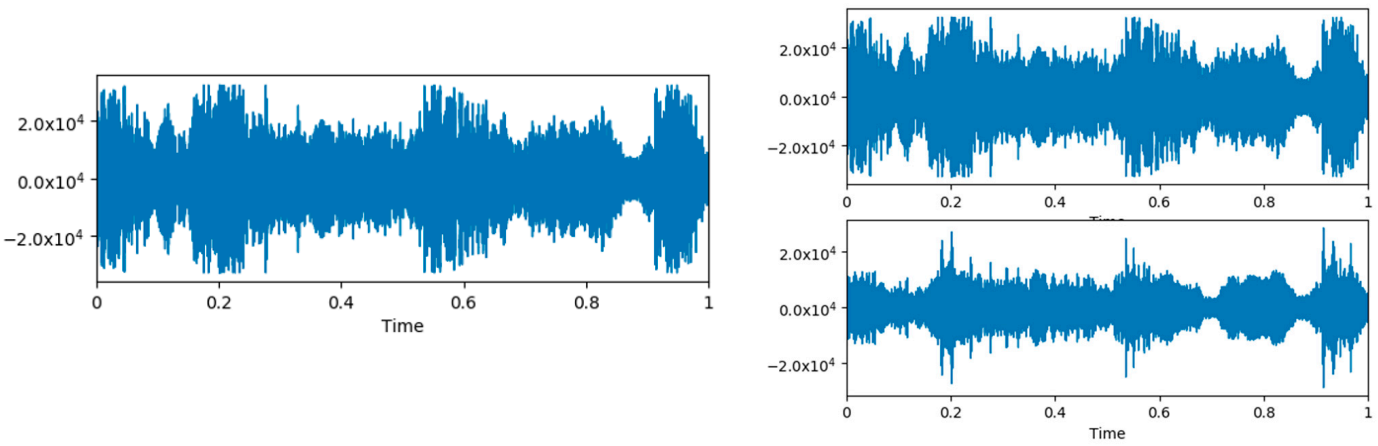
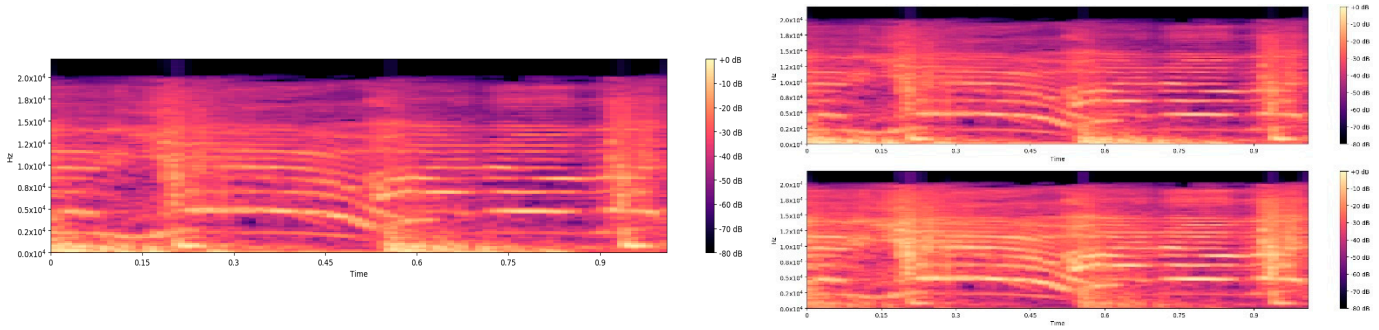


Figure 1. Audio waveforms. (**a**) Real mono audio; (**b**) fake stereo audio.

The rest of this paper is organized as follows. In Section 2, the detail of the fake stereo corpus is given. Section 3 describes an identification algorithm based on MFCC features and SVM classification. Section 4 describes an end-to-end deep algorithm for identifying fake stereo audio. In Section 5, we present the experimental results in various cases. Finally, the conclusions are given in Section 6.



(**a**) (**b**)

Figure 2. Audio spectrograms. (**a**) Real mono audio; (**b**) fake stereo audio.

2. Stereo Faking Corpus

To evaluate the performance of the proposed identification algorithm, a new fake stereo audio corpus was established as described below.

2.1. Stereo Faking

The aim of stereo faking is to convert mono audio into an artificial stereo. There are two typical methods to create fake stereo audio. One is called Channel Copy. First, a faked channel can be created by coping the original channel of the mono audio. Then, by integrating these two channels, the fake stereo audio will be obtained by Equation (1):

$$Y = \{x_l, \hat{x}_l\} = \begin{cases} x_l, & left\ channel \\ \hat{x}_l, & right\ channel \end{cases}, \tag{1}$$

where $\hat{x}_l = x_l, l = 1, 2, 3, \ldots L$, x_l is the mono audio channel, $\hat{x}_l$ is the fake audio channel, L is the length of the audio, and Y represents fake stereo audio. It is easy to identify the fake stereo created by Channel Copy because the sample values of its two channels are always the same.

The Haas effect [29] is another method to create a more deceptive fake stereo audio. The Haas effect is a phenomenon where human ears perceive two sounds as being one. When we hear two sounds within 40 milliseconds of one another, human ears interpret them as being the same sound. Haas has shown through experiments that if the time difference between two sound waves from the same sound source reaching the listener is within 5~35 ms, one cannot distinguish between the two sound sources. If the delay time is 35~50 ms, the human ear begins to perceive the two sound sources. If the time difference is greater than 50 ms, the human ear can distinguish the position of the two sound sources. Haas' description of how the different delays of the two sound sources reflect to the human ear is called the Haas effect. This effect helps to establish a stereo listening environment. In the specific implementation, we adopt a high-pass filter to simulate this delay effect. Additionally, the Haas effect can be expressed by Equations (2) and (3):

$$\hat{x}_l = F(x_l), l = 1,\ 2, 3, \ldots, L, \tag{2}$$

$$Y = \{x_l, \hat{x}_l\} = \begin{cases} x_l, & left\ channel \\ \hat{x}_l, & right\ channel \end{cases}, \tag{3}$$

where $\hat{x}_l \neq x_l,\ l = 1, 2, 3, \ldots L$, and F denotes the high-pass filter to deal with the forged channel. A high-pass filter is a filter that allows frequencies higher than a certain cut-off frequency to pass while greatly attenuating lower frequencies. Unnecessary low-frequency

components or low-frequency interference is removed from the signal. The high-pass filter F in this work can be defined as Equation (4):

$$\hat{x}_l = \frac{1}{1 + 2\pi f_{cut-off} T}(x_l - x_{l-1} + \hat{x}_{l-1}), \quad (4)$$

where T denotes the cycle and f_{cof} is the cut-off frequency (cof).

2.2. Real and Fake Stereo Datasets

The real stereo audio in the corpus comes from the following two resources: IMDbTop250 [30] movies and QQ Music [31] songs. IMDb (Internet Movie Database) is an online database about movie actors, movies, TV shows, TV stars and movie productions. Ten movies in the IMDb Top250, which are Shawshank's Redemption, Hachiko, Jurassic Park, Seven Deadly Sins, Dream Stealing Space, Truman, Forrest Gump, Return of Tarzan, Detective Chinatown 2, and Number One Player, are considered in this work. First, the soundtracks of these movies were extracted, and then each of them was segmented into 30 min. Therefore, the first real stereo dataset consisted of a total 5 h of audio clips, which was named FILM. QQ Music is a music streaming platform in China. It provides tens of millions of songs. A total of 91 songs in QQ Music were downloaded as the second real stereo dataset, named MUSIC.

Each clip in the real stereo datasets was segmented into 1 s and used to create a fake stereo audio by the Haas effect mentioned in Section 2.1. Finally, we constructed a corpus for fake stereo audio identification. To ensure the validity and fairness, the corpus was randomly divided into training and testing datasets, which is shown in Table 1. The samples in the training datasets were used to train the model, and the samples in the testing datasets were used to test the performance of the model. It should be noted that the samples in the testing datasets were not included in the training datasets. The ratio of training datasets to testing datasets was about 7:3.

Table 1. Information of two major datasets.

Corpus	Training Datasets	Testing Datasets
FILM	12,600	5400
MUSIC	16,684	5895

According to the description in Section 1, the cut-off frequency of the high-pass filter is a critical parameter in the Haas effect. On the other hand, since the counterfeiter may forge any channel in the stereo, the two channels in the stereo to be forged separately, and five kinds of cut-off frequencies in the high-pass filter are considered in order to comprehensively evaluate the performance of the algorithm. Specially, the cut-off frequency for the training set is fixed at 200 Hz, and the cut-off frequencies for the testing set are configured to 200 Hz, 400 Hz, 600 Hz, 800 Hz, and 1000 Hz, respectively. The format of each sample is WAV, a 44,100 Hz sampling rate, and 16-bit quantization.

3. MFCC Features and SVM Classification Algorithms

In this section, we first introduce the extraction of MFCC features. Then, the SVM classifier used as classification is briefly described. Finally, the proposed identification algorithm of fake stereo audio is presented, which is based on MFCC features and SVM classification.

3.1. MFCC Feature Extraction

Mel-frequency cepstral coefficients (MFCCs) are widely used in speech recognition functions. They were proposed by Davis and Mermelstein in the 1980s [32] and continue to be one of the most advanced technologies thereafter. Mel-frequency cepstrum is a linear transformation of the logarithmic energy spectrum based on the non-linear Mel scale of

sound frequency. Mel-frequency cepstral coefficients (MFCCs) are the coefficients that make up the Mel-frequency cepstral coefficients. They are derived from the cepstrum of audio fragments. The difference between the cepstrum and Mel-frequency cepstrum is that the band division of the Mel-frequency cepstrum is equally spaced on the Mel scale, which is more approximate than the linearly spaced frequency band used in the normal cepstrum, the human auditory system. Such a non-linear representation can provide a better representation of the sound signal in multiple fields. Chowdhury et al. [33] provided a fusion of MFCC and Linear Predictive Coding (LPC) for speaker recognition. Sahidullah et al. [34] proposed a novel family of windowing techniques to compute MFCCs for automatic speaker recognition from speech.

The MFCC models the spectral energy distribution in a perceptually meaningful way. Before calculating the MFCC, in order to amplify the high-frequency components, a finite impulse response (FIR) high-pass filter should be used to pre-emphasize the input speech $X = [x(1), x(2), \ldots, x(l), \ldots x(L)]$ as follows:

$$x'(i) = x(i) - 0.95x(i-1), \tag{5}$$

where L is the size of the input audio.

Then, by multiplying each frame by the Hamming window, the emphasis signal $x'(i)$ is divided into overlapped frames. The Hamming window is:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right),\ 0 \le n \le N-1, \tag{6}$$

where N is the length of the window.

Next, the frequency spectrum is obtained by applying Fast Fourier Transform (FFT) on each windowed frame. Then, the spectrum is decomposed into multiple sub-bands using a set of triangular Mel-scale bandpass filters. Let $E(b)$, $0 \le b \le B$ represent the sum of the power spectral coefficients in the bth sub-bands, where B is the total number of filters.

The MFCC can be calculated by applying the discrete cosine transform (DCT) to the logarithm of $E(b)$ as follows:

$$C(s) = \sum_{b=0}^{B-1} \log_{10}(1 + E(b)) \cos\left(s\frac{\pi}{B}(b+0.5)\right),\ 0 \le s \le S, \tag{7}$$

where S is the length of the MFCC. In this work, we extracted the MFCC features of the left and right channels of the stereo audio separately, and stitched the features of the right channel to the left channel, and finally we obtained $2S$ dimensional features.

In Figure 3, the boxplots of the MFCC features from the real and fake stereo audios are shown. Various cut-off frequencies of high-pass filtering are considered. It can be seen that the distributions of the MFCC features in FILM and MUSIC datasets have an obvious change between the real and fake stereo audios. Since the MFCC features can expose the effect caused by stereo faking, they are used to identify the fake stereo.

3.2. SVM Classification

An SVM [35] is a supervised learning technique that is a powerful discriminative classifier and it can be used in fake stereo audio detection. As shown in Figure 4, an SVM is a binary classifier which models the decision boundary between two classes as a separating hyperplane. In fake stereo audio detection, one class is made up of the MFCC features from real stereo audio (labeled as +1), and the other class comprises the features from the fake one (labeled as -1). With the labeled training datasets, the trained SVM classifier finds a separating hyperplane to maximize the d of separation between these two classes More detailed information about the SVM can be found in [35].

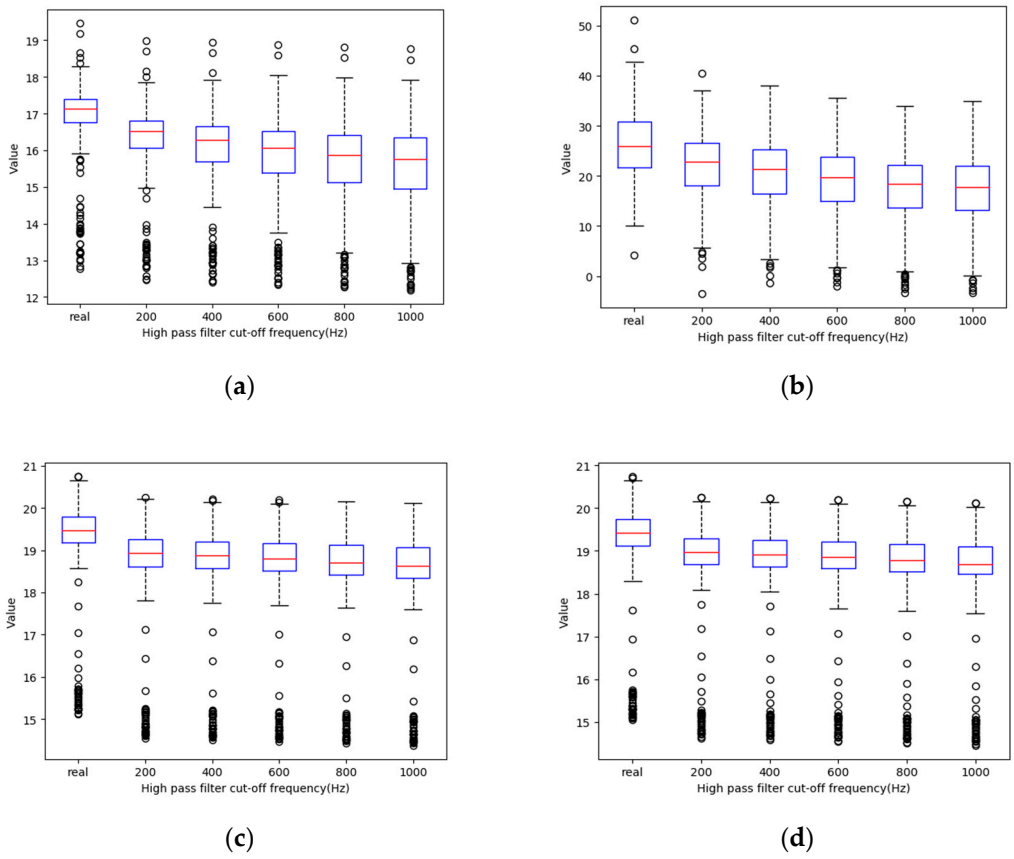


Figure 3. Distributions of the MFCC components for the real and fake stereo audios (real denotes the real stereo audio, 200 to 1000 denotes the fake stereo audio ($f_{cut-off}$ are 200 Hz to 1000 Hz, respectively)). (**a**) First component of right channel in FILM (**b**) 2nd component of left channel in FILM (**c**) 1st component of right channel in MUSIC (**d**) 2nd component of left channel in MUSIC.

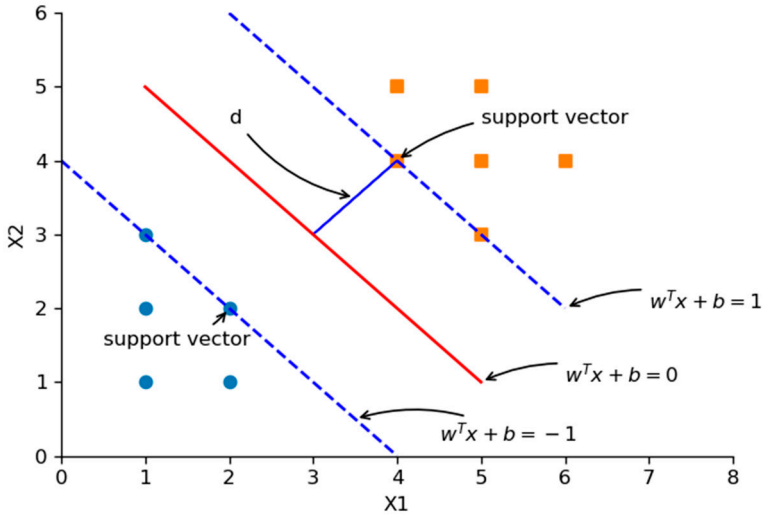


Figure 4. Principle of support vector machine (SVM).

The classification decision function of an SVM can be defined as:

$$f(c) = sign\left(\sum_{i=1}^{N} \alpha_i y_i K(c_i, c_j) + b\right), \quad (8)$$

where y_i are the real output values, $\sum_{i=1}^{N} \alpha_i y_i = 0$ and $\alpha_i > 0$. The support vectors c_j, their corresponding weights α_i and bias b are calculated on the training set. The kernel function $K(c_i, c_j)$ can be defined as $K(c, y) = \langle c, y \rangle$; this is actually the inner product of the original input space. The kernel function is to replace the inner product of the feature space with the inner product of the input vector, to achieve the purpose of mapping the input space data to a high-dimensional space. In a high-dimensional space, the two classes are easier to separate with a hyperplane. Finally, the sigmoid (*sign*) activation function is used to obtain the final classification result.

3.3. Algorithm

A real and fake stereo based on the left channel in Figure 5 means that the real mono audio is used as the left channel of fake stereo audio, the right channel is faked, and vice versa. Considering that forensics do not know which channel the forger will use for forgery, it is necessary to train the left and right channels separately. Therefore, we used two SVM models for each dataset, called the first SVM classifier and the second SVM classifier.

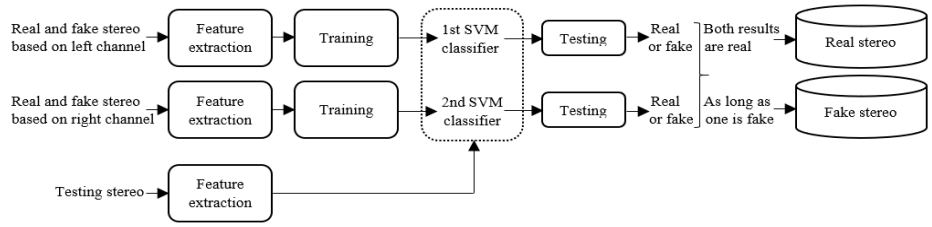


Figure 5. MFCC features and SVM classification algorithm flowchart.

Our algorithm is divided into the training and testing stages. In the training stage, the training dataset is composed of a real stereo dataset and two fake stereo datasets. MFCC coefficients are extracted from the training set as the detection features. The features are used as the training features to train two SVM classifiers. Additionally, each SVM classifier can be used to identify whether the testing stereo audio is faked or not. In the testing stage, the MFCC features from the real and fake stereo audio in the test dataset are extracted. Then, they are used as the input into the trained SVM classifiers. These two classification results are combined to obtain a final detection result. If both results are real, the testing stereo audio is identified as real. If at least one result is fake, the testing stereo audio is identified as fake.

4. End-to-End Deep Convolutional Neural Network

In this section, we give a general description of the end-to-end deep CNN model. Then, the Loss function is briefly described. Finally, the proposed identification algorithm of fake stereo audio is presented.

4.1. Network Framework

Deep learning is a branch in the field of machine learning. It has achieved many results in search technology, data mining, machine learning, machine translation, natural language processing, multimedia, speech, recommendation, personalization technology, and other related fields. It does not need to extract features in advance, because its convolutional layer can automatically extract features. The convolutional neural network is a classic model in deep learning. Lecun et al. [36] proposed LeNet-5 and achieved good results in MNIST digital handwriting recognition. In this work, we designed a CNN with a five-layer convolutional layer and a pooling layer. The convolutional layer is responsible for extracting acoustic features, and the pooling layer is responsible for reducing the dimensionality of the output of the previous convolutional layer and speeding up the convergence.

The end-to-end deep model D is designed to distinguish between real and fake stereo audio. As shown in Figure 6, the model consists of five convolution groups and three

fully connected layers: each group includes a convolutional layer with 64 filters, followed by a LeakyReLU activation and a Max pooling (the pooling layer of the last group of convolutional groups uses average pooling, because this can effectively prevent overfitting). To control the size of downsampling, the size of the convolution kernel of each layer and the pooling window of the pooling layer are slightly different. The specific parameters are shown in Figure 5. Finally, fully connected layers coupled with a sigmoid layer are commonly used to generate an output class probability. A Dropout layer is added between the fully connected layers, and p is 0.5, which can also effectively prevent overfitting.

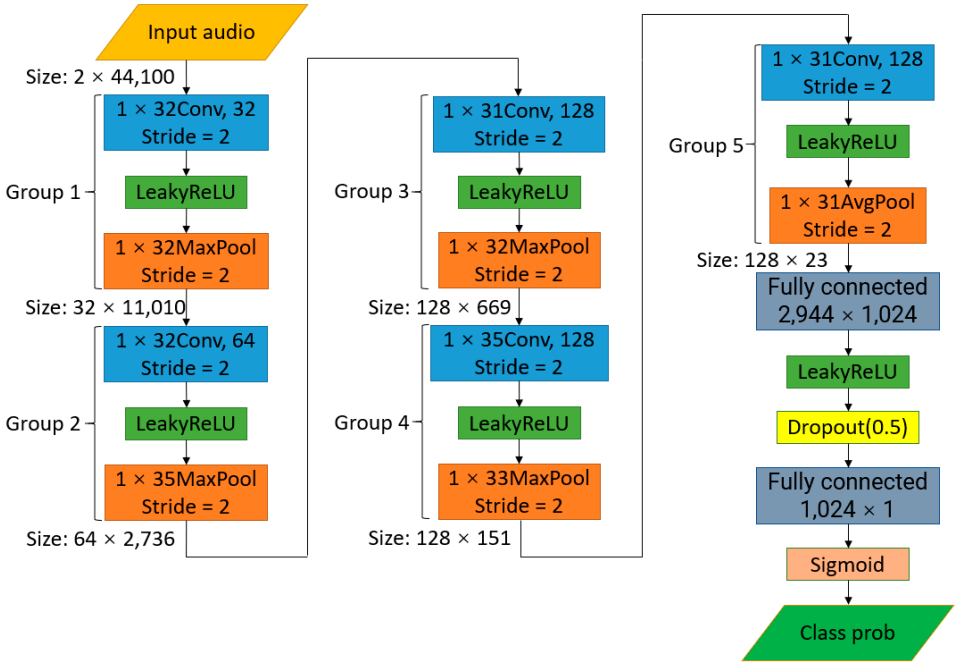


Figure 6. Architecture of deep end-to-end model. Stereo audio will first pass through five convolutional layers (group 1 to group 5—each convolutional layer contains 64 convolution kernels, LeakyReLU activation function and Max pooling, and the last layer uses average pooling). Then, the discriminant probability will be output through the three-layer fully connected layer and the sigmoid activation function. Size represents the size of the sample before it passes through each layer of the network.

4.2. Loss Function

The purpose of the model D is to distinguish between real and fake stereo audio accurately. Hence, the loss function of the deep end-to-end model D can be defined as:

$$L_d = L_\alpha + L_\beta, \tag{9}$$

where L_α represents to recognize real stereo audio as real, and L_β denotes that the fake stereo audio is recognized as fake. The Binary Cross Entropy loss function (BCELoss) is adopted in this work:

$$L_\alpha = t_1 log D(X) + (1 - t_1) log(1 - D(X)), \tag{10}$$

$$L_\beta = t_2 log D(Y) + (1 - t_2) log(1 - D(Y)), \tag{11}$$

where $D(\cdot)$ denotes the output of the deep end-to-end model, t represents the label output by the deep end-to-end model D; in this work, t_1 is set to 1, and t_2 is set to 0. X denotes the real stereo audio and the Y represents the fake stereo audio generated by the Hass effect.

4.3. Algorithm

Figure 7 is a flow chart of the proposed CNN algorithm. Similar to the algorithm in Section 3, in the training stage, we also train two CNN models for two channels. In the testing stage, we directly feed the test stereo audio to the two trained CNN models to obtain the classification probability and combine the results of the two CNN models to distinguish the authenticity of the testing stereo audio. For more details, see Figure 6 or refer to Section 3.3.

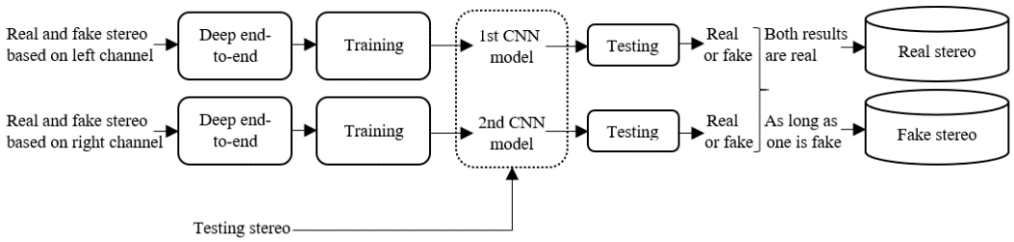


Figure 7. Deep end-to-end algorithm flowchart. In the training stage, the dataset for training is composed of a real stereo set and two fake stereo sets (1st CNN classifier and 2nd CNN classifier). They are used as the training features to train two CNN classifiers. Additionally, each CNN classifier can be used to identify whether a testing stereo audio is faked or not. In the testing stage, the testing datasets are used as the input into the trained CNN classifiers. These two classification results are combined to obtain a final detection result. If both results are real, the testing stereo audio is identified as real. If at least one result is fake, the testing stereo audio is identified as fake.

5. Experimental Results

In this section, we present the experimental results of the proposed method. First, the experimental setup is described. Then, the experimental results for intra-dataset and cross-dataset are given to evaluate the performance of the proposed method. Accuracy rate (ACC), false acceptance rate (FAR) and false reject rate (FRR), are adopted as evaluation metrics of our model. Specifically, the negative sample in this work is fake stereo audio and the positive sample is real stereo audio. Therefore, the FAR means the ratio of the number of negative samples incorrectly identified as positive samples to the total number of negative samples, and the FRR is the ratio of the number of positive samples incorrectly identified as negative samples to the total number of positive samples.

5.1. Experimental Setup

In our experiment, the fake stereo audio is created by the Haas effect. For the algorithm based on MFCC features and SVM classification, each dataset in the corpus is divided into two disjoint parts by the ratio of 6:4. The first 60% is the training set, and the remaining 40% is used for verification. For the extraction of its MFCC features, S is 40. For the SVM classifier, the penalty coefficient C is set to 0.4, and the linear kernel function is adopted. For the end-to-end deep algorithms, we implement our models using PyTorch [37] and used Nvidia GeForce GTX 1080Ti GPU with 3584 CUDA cores for training our models. The input dimensions of our network are $1 \times 44{,}100$ and the output dimension is 1×1. The learning rate is set to 0.00001, and the optimization algorithm uses adaptive moment estimation (Adam). The training epoch is 10 rounds.

5.2. Intra-Dataset Evaluation

This evaluation case means that the dataset for testing is the same as that for training. In this experiment, our detection ACC can reach up to 99%. Table 2 shows the identification performance of the proposed method. It can be seen that in the FILM dataset, the FAR of the CNN model is slightly higher than that of the SVM model, and the FRR is almost close to 0. The FAR of the CNN model in the MUSIC dataset reaches 0, which is significantly better than the SVM, and the FRR is maintained at around 1%. Hence, the experimental results indicate that the method's performance is remarkable.

Table 2. FAR and FRR for intra-dataset evaluation (%).

	FILM-SVM	FILM-CNN	MUSIC-SVM	MUSIC-CNN
1st	0.02/0.00	0.04/0.00	0.08/1.42	0.00/1.09
2nd	0.06/0.02	0.09/0.00	0.12/0.92	0.00/0.75

5.3. Various High-Pass Filter Parameters for Testing

In this experiment, two models are trained with left and right channels at the cut-off frequency 200 Hz. The stereo audio faked with various cut-off frequencies are tested. Since the positive samples are the same in this experiment, the FRR is consistent with Section 5.2. Hence, we mainly focus on the ACC and FAR. The experimental results for left and right channels are shown in Tables 3 and 4, respectively. It can be found that the accuracy of the two algorithms also reached 99%. For the first model, in the FILM dataset, the FAR of the CNN model is slightly higher than that of the SVM model, while the FAR of the SVM model and the CNN model in the MUSIC dataset are both close to 0. For the second model, regardless of FILM or MUSIC, its performance is maintained at a high level, in which the ACC remains at 99%, and the FAR tends to 0. Meanwhile, it can be seen that the performance can be kept at a high and stable level with the increase in cut-off frequency. This shows that our proposed algorithm has a good robustness.

Table 3. ACC and FAR of the 1st model with various f_{cof} (%).

Cut-Off Frequency (Hz)	FILM-SVM	FILM-CNN	MUSIC-SVM	MUSIC-CNN
200	99/0.02	99/0.04	99/0.08	99/0.00
400	100/0.00	99/0.06	99/0.00	99/0.00
600	100/0.00	99/0.44	99/0.02	99/0.00
800	100/0.00	99/1.69	99/0.02	99/0.00
1000	100/0.00	99/2.74	99/0.00	99/0.00

Table 4. ACC and FAR of the 2nd model with various f_{cof} (%).

Cut-Off Frequency (Hz)	FILM-SVM	FILM-CNN	MUSIC-SVM	MUSIC-CNN
200	99/0.02	99/0.09	99/0.08	99/0.00
400	99/0.00	99/0.04	99/0.00	99/0.00
600	99/0.00	99/0.04	99/0.00	99/0.00
800	99/0.00	99/0.04	99/0.00	99/0.00
1000	99/0.06	99/0.09	99/0.00	99/0.00

5.4. Cross-Dataset Evaluation

In practical forensic scenarios, the channels of suspected audio are variable in recording devices and environments. In our corpus, FILM is movie recording, and MUSIC is songs. Their recording equipment and environment are also different from each other. Hence, cross-database evaluation is necessary. Similar to Sections 5.2 and 5.3, when using one of training dataset for training, the stereo from all testing datasets is tested. Note that the datasets used for training and testing are different. In view of the fact that the FRR is an evaluation index for positive samples, it cannot reflect the robustness of the model to Hass effect forgery methods. Therefore, we focus on the two indicators of the ACC and FAR in the cross-dataset evaluation.

The MFCC features and SVM classification algorithm detection performance of the cross-dataset evaluation is shown in Tables 5 and 6. Compared with Table 2, the cross-dataset performance is a little worse than the intra-dataset case. In particular, the model trained in the FILM training dataset has an accuracy rate of less than 60% when detecting the MUSIC testing dataset, and the FAR is higher than 0.8. Only the model trained in

the MUSIC training dataset shows strong robustness in all two testing datasets. The FRR fluctuates between 0% and 4%. This shows that at the MFCC feature level, the samples from the MUSIC datasets are more difficult to distinguish than the samples from the FILM datasets. Therefore, the SVM classifier trained on the FILM training dataset is not robust enough to effectively detect the MUSIC testing dataset.

Table 5. ACC, FAR and FRR for the 1st SVM classifier (%).

	FILM	**MUSIC**
FILM	99/0.02/0.00	55/88.00/0.75
MUSIC	99/0.00/0.30	99/0.08/1.42

Table 6. ACC, FAR and FRR for the 2nd SVM classifier (%).

	FILM	**MUSIC**
FILM	99/0.06/0.02	54/86.00/3.74
MUSIC	99/0.00/0.43	99/0.12/0.92

The end-to-end deep algorithm detection performance of the cross-dataset evaluation is shown in Tables 7 and 8. It can be seen that the recognition accuracy rate has reached more than 99%, and the FAR is also below 1%, even reaching 0%; the FRR is maintained at around 1%. It can be found that the robustness of the deep end-to-end algorithm is significantly better than that of the MFCC features and SVM classification algorithm. Additionally, the convergence speed of the deep end-to-end algorithm is less than that of the MFCC features and SVM classification algorithm. The reason why the CNN model is so robust is that the model does not need to extract handcrafted features in advance. The CNN model is trained to focus on the relationship between the two stereo audio channels. The features extracted by our five-layer convolutional layer also mainly reflect this relationship. The stereo audio forged by the Haas effect is produced through a series of operations on mono audio. Compared with the real stereo audio, the fake stereo audio produced by the Haas effect does not show this relationship between the two channels. In other words, the MFCC pays more attention to the characteristics of audio content, while our CNN model pays more attention to the characteristics of the relationship between stereo audio channels, while FILM and MUSIC are two completely different datasets in content. Therefore, the CNN model that mainly focuses on the relationship between the two channels has good robustness.

Table 7. ACC, FAR and FRR for the 1st CNN model (%).

	FILM	**MUSIC**
FILM	99/0.04/0.00	99/0.00/0.83
MUSIC	99/0.00/0.22	99/0.00/1.09

Table 8. ACC, FAR and FRR for the 2nd CNN model (%).

	FILM	**MUSIC**
FILM	99/0.09/0.00	99/0.87/0.88
MUSIC	99/0.04/0.24	99/0.00/0.75

6. Conclusions

In this work, we proposed two algorithms for identifying fake stereo audio (1) MFCC features and SVM classification algorithm, and (2) end-to-end deep algorithm. The experimental results are given in the form of comparison between the two algorithms. After

rigorous experiments, the results show that both of our proposed algorithms can effectively detect fake stereo audio. The recognition accuracy can reach 99%, and the minimum false acceptation rate can reach 0%. Additionally, the deep end-to-end algorithm is significantly better than the MFCC features and SVM classification algorithm in robustness. Combining all the experimental results, the end-to-end deep algorithm is more suitable for detecting fake stereo audio.

The possible practical use of this work is mainly for music websites, apps and their users. When these websites and apps collect some songs, they can use our model to detect whether the songs are faked or not. Similarly, when users download their favorite songs from music websites and apps, they can also use our model for fake stereo audio detection. In the future, our work has two goals: one is to improve the robustness, security and reliability of our forensics model, and the other is to conduct anti-forensics research on the forensics model.

Author Contributions: Conceptualization, D.Y.; formal analysis, T.L.; funding acquisition, D.Y. and R.W.; investigation, T.L.; methodology, T.L.; project administration, D.Y.; supervision, D.Y.; validation, T.L. and N.Y.; writing—original draft, T.L.; writing—review and editing, T.L., D.Y., R.W., N.Y. and G.C. All authors will be informed about each step of manuscript processing including submission, revision, revision reminder, etc. via emails from our system or assigned Assistant Editor. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 61300055), Zhejiang Natural Science Foundation (Grant No. LY20F020010, LY17F020010), Ningbo Natural Science Foundation (Grant No. 202003N4089), Ningbo Science and Technology Innovation 2025 Major Project (Grant No. 2018B10010, 2019B10075).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available on request due to privacy. The data presented in this study are available on request from the corresponding author. Since the data are private, they are not publicly available.

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

References

1. Yongqiang, B.; Ruiyu, L.; Yun, C.; Chonghong, G.; Qinyun, W. Research Progress on Key Technologies of Audio Forensics. *J. Data Acquis. Process.* **2016**, *31*, 252–259.
2. Luo, D.; Yang, R.; Li, B.; Huang, J. Detection of Double Compressed AMR Audio using Stacked Autoencoder. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 432–444. [CrossRef]
3. Luo, D.; Yang, R.; Huang, J. Detecting Double Compressed AMR Audio using Deep Learning. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 4–9 May 2014; pp. 432–444.
4. Wu, H.; Wang, Y.; Huang, J. Identification of Electronic Disguised Voices. *IEEE Trans. Inf. Forensics Secur.* **2014**, *9*, 489–500. [CrossRef]
5. Wu, H.; Wang, Y.; Huang, J. Blind detection of Electronic Disguised Voice. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 3013–3017.
6. Xu, H.; Yan, D.; Yang, F.; Wang, R.; Jin, C.; Xiang, L. Detection algorithm of Electronic Disguised Voice based on Convolutional Neural Network. *Telecommun. Sci.* **2018**, *34*, 46–57.
7. Luo, D.; Korus, P.; Huang, J. Band Energy difference for source attribution in Audio Forensics. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2179–2189. [CrossRef]
8. Zou, L.; He, Q.; Feng, X. Cell Phone Verification from Speech Recordings using Sparse Representation. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, South Brisbane, QLD, Australia, 19–24 April 2015; pp. 1787–1791.
9. Qi, S.; Huang, Z.; Li, Y.; Shi, S. Audio Recording Device Identification based on Deep Learning. In Proceedings of the International Conference on Signal & Image Processing, Beijing, China, 13–15 August 2016; pp. 426–431.
10. Gałka, J.; Grzywacz, M.; Samborski, R. Playback Attack Detection for Text-Dependent Speaker Verification over Telephone Channels. *Speech Commun.* **2015**, *67*, 143–153. [CrossRef]

11. Wu, T.; Yan, D.; Li, X.; Wang, R. Detection of Operation Type and Order for Digital Speech. In Proceedings of the 7th Conference on Sound and MUSIC Technology (CSMT); Lecture Notes in Electrical Engineering. Springer: Singapore, 2020; Volume 635, pp. 25–37.
12. Yang, R.; Shi, Y.; Huang, J. Defeating Fake-Quality MP3. In Proceedings of the 11th ACM Workshop on Multimedia and Security, MM & Sec, Princeton, NJ, USA, 7–8 September 2009; Volume 9, pp. 117–124.
13. Mascia, M.; Canclini, A.; Antonacci, F.; Tagliasacchi, M.; Sarti, A.; Tubaro, S. Forensic and anti-forensic analysis of indoor/outdoor classifiers based on acoustic clues. In Proceedings of the 2015 23rd European Signal Processing Conference, Nice, France, 31 August–4 September 2015; pp. 2072–2076.
14. Vijayasenan, D.; Kalluri, S.B.; K, S.; Issac, A. Study of Wireless Channel Effects on Audio Forensics. In Proceedings of the 2016 22nd Annual International Conference on Advanced Computing and Communication, Bangalore, India, 8–10 September 2016; pp. 33–37.
15. Hadoltikar, V.A.; Ratnaparkhe, V.R.; Kumar, R. Optimization of MFCC parameters for mobile phone recognition from audio recordings. In Proceedings of the 2019 3rd International conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 12–14 June 2019; pp. 777–780.
16. Zhao, H.; Malik, H. Audio forensics using acoustic environment traces. In Proceedings of the 2012 IEEE Statistical Signal Processing Workshop, Coimbatore, India, 12–14 June 2019.
17. Jiang, Y.; Leung, F.H.F. Mobile phone identification from speech recordings using Weighted Support Vector Machine. In Proceedings of the IECON 2016—42nd Annual Conference of the IEEE Industrial Electronics Society, Florence, Italy, 23–26 October 2016; pp. 963–968.
18. Mitra, V.; Franco, H.; Graciarena, M.; Vergyri, D. Medium-duration modulation cepstral feature for robust speech recognition. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 4–9 May 2014; pp. 1749–1753.
19. Athanaselis, T.; Bakamidis, S.; Giannopoulos, G.; Dologlou, I.; Fotinea, E. Robust speech recognition in the presence of noise using medical data. In Proceedings of the 2008 IEEE International Workshop on Imaging Systems and Techniques, Chania, Greece, 10–12 September 2008; pp. 349–352.
20. Subramanian, A.S.; Weng, C.; Yu, M.; Zhang, S.; Xu, Y.; Watanabe, S.; Yu, D. Far-Field Location Guided Target Speech Extraction Using End-to-End Speech Recognition Objectives. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 7299–7303.
21. Fan, C.; Yi, J.; Tao, J.; Tian, Z.; Liu, B.; Wen, Z. Gated Recurrent Fusion With Joint Training Framework for Robust End-to-End Speech Recognition. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*; IEEE: New York, NJ, USA, 2021; Volume 29, pp. 198–209.
22. Toruk, M.M.; Gokay, R. Short Utterance Speaker Recognition Using Time-Delay Neural Network. In Proceedings of the 2019 16th International Multi-Conference on Systems, Signals & Devices, Istanbul, Turkey, 21–24 March 2019; pp. 383–386.
23. Huang, C. Exploring Effective Data Augmentation with TDNN-LSTM Neural Network Embedding for Speaker Recognition. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop, Singapore, 14–18 December 2019; pp. 291–295.
24. Jagiasi, R.; Ghosalkar, S.; Kulal, P.; Bharambe, A. CNN based speaker recognition in language and text-independent small scale system. In Proceedings of the 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), Palladam, India, 12–14 December 2019; pp. 176–179.
25. Desai, N.; Tahilramani, N. Digital Speech Watermarking for Authenticity of Speaker in Speaker Recognition System. In Proceedings of the 2016 International Conference on Micro-Electronics and Telecommunication Engineering, Ghaziabad, India, 22–23 September 2016; pp. 105–109.
26. Reis, P.M.G.I.; Lustosa da Costa, J.P.C.; Miranda, R.K.; Del Galdo, G. ESPRIT-Hilbert-Based Audio Tampering Detection With SVM Classifier for Forensic Analysis via Electrical Network Frequency. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 853–864. [CrossRef]
27. Lin, X.; Liu, J.; Kang, X. Audio Recapture Detection with Convolutional Neural Networks. *IEEE Trans. Multimed.* **2016**, *18*, 1480–1487. [CrossRef]
28. Ciaburro, G. Sound Event Detection in Underground Parking Garage Using Convolutional Neural Network. *Big Data Cogn. Comput.* **2020**, *4*, 20. [CrossRef]
29. Haas, H. The influence of a single echo on the audibility of speech. *J. Audio Eng. Soc.* **1972**, *20*, 146–159.
30. IMDbTop250. Available online: https://www.imdb.com/chart/top (accessed on 28 June 2021).
31. QQ Music. Available online: https://music.qq.com (accessed on 28 June 2021).
32. Davis, S.; Mermelstein, P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 357–366. [CrossRef]
33. Chowdhury, A.; Ross, A. Fusing MFCC and LPC Features Using 1D Triplet CNN for Speaker Recognition in Severely Degraded Audio Signals. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 1616–1629. [CrossRef]
34. Sahidullah, M.; Saha, G. A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition. *IEEE Signal Process. Lett.* **2013**, *20*, 149–152. [CrossRef]
35. Bishop, C. *Pattern Recognition and Machine Learning*; Springer Science+Business Media: New York, NY, USA, 2006.

36. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
37. Collobert, R.; Koray, K.; Farabet, C. Torch7: A Matlab-like Environment for Machine Learning. 2011. Available online: http://publications.idiap.ch/downloads/papers/2011/Collobert_NIPSWORKSHOP_2011.pdf (accessed on 28 June 2021).