information

MDPI

Article

Source Cell-Phone Identification in the Presence of Additive Noise from CQT Domain

Tianyun Qin, Rangding Wang *, Diqun Yan and Lang Lin

College of Information Science and Engineering of Ningbo University, Ningbo 315211, China; qintianyun111@163.com (T.Q.); yandiqun@nbu.edu.cn (D.Y.); ll_linlang@163.com (L.L.)

* Correspondence: wangrangding@nbu.edu.cn; Tel.: +86-138-0589-8088

Received: 18 July 2018; Accepted: 13 August 2018; Published: 17 August 2018

Abstract: With the widespread availability of cell-phone recording devices, source cell-phone identification has become a hot topic in multimedia forensics. At present, the research on the source cell-phone identification in clean conditions has achieved good results, but that in noisy environments is not ideal. This paper proposes a novel source cell-phone identification system suitable for both clean and noisy environments using spectral distribution features of constant Q transform (CQT) domain and multi-scene training method. Based on the analysis, it is found that the identification difficulty lies in different models of cell-phones of the same brand, and their tiny differences are mainly in the middle and low frequency bands. Therefore, this paper extracts spectral distribution features from the CQT domain, which has a higher frequency resolution in the mid-low frequency. To evaluate the effectiveness of the proposed feature, four classification techniques of Support Vector Machine (SVM), Random Forest (RF), Convolutional Neural Network (CNN) and Recurrent Neuron Network-Long Short-Term Memory Neural Network (RNN-BLSTM) are used to identify the source recording device. Experimental results show that the features proposed in this paper have superior performance. Compared with Mel frequency cepstral coefficient (MFCC) and linear frequency cepstral coefficient (LFCC), it enhances the accuracy of cell-phones within the same brand, whether the speech to be tested comprises clean speech files or noisy speech files. In addition, the CNN classification effect is outstanding. In terms of models, the model is established by the multi-scene training method, which improves the distinguishing ability of the model in the noisy environment than single-scenario training method. The average accuracy rate in CNN for clean speech files on the CKC speech database (CKC-SD) and TIMIT Recaptured Database (TIMIT-RD) databases increased from 95.47% and 97.89% to 97.08% and 99.29%, respectively. For noisy speech files with seen noisy types and unseen noisy types, the performance was greatly improved, and most of the recognition rates exceeded 90%. Therefore, the source identification system in this paper is robust to noise.

Keywords: source cell-phone identification; additive noise; CQT; CNN; multi-scene training; noise robustness

1. Introduction

With the development and advancement of digital multimedia and Internet technologies, a variety of powerful and easy-to-operate digital media editing software has emerged, bringing new problems and challenges to the availability of collected data–multimedia security issues. Recording device identification is a branch of multimedia forensics technology, and has research significance. Compared with recorders, cameras, DVs, etc., mobile phones are more popular and convenient. More and more people are using mobile phones to collect the scenes they hear, and even use the recording file as

evidence before courts or other law enforcement agencies. Therefore, source cell-phone identification is a hot topic for many forensic researchers.

In recent years, source cell-phone identification has achieved great research results. In the beginning, many researchers used cepstral coefficients or features based on cepstral coefficients as the fingerprint of the device. C. Hanilci et al. [1] extracted the Mel frequency cepstral coefficient (MFCC) from the recording file as a device-distinguishing feature, and 14 different models of cell-phones were evaluated in the experiment. The closed-collection recognition rate reached 96.42% using SVM classifiers. In a follow-up study, C. Hanilci et al. [2] used SVM to compare MFCC, linear frequency cepstral coefficient (LFCC), Bark frequency cepstral coefficient (BFCC) and linear predictive cepstral coefficient (LPCC). Their comparison covered various kinds of feature optimization, including feature normalization, cepstral mean normalization, cepstral mean and variance normalization, and delta and double-delta coefficients. The experimental results showed that while baseline MFCCs outperformed other types of features, work of both cepstral mean and variance normalization yielded superior performance for LPCCs (with only slightly better results than MFCCs). In addition, C. Kotropoulos et al. [3] extracted MFCC from any recorded speech signal at a frame level. The MFCC from each recording device trained a Gaussian Mixture Model (GMM) with diagonal covariance matrices. A Gaussian super vector (GSV) is derived by concatenating the mean vectors and the main diagonals of the covariance matrices that is used as a template for each device. The best identification accuracy (97.6%) was obtained by the Radial Basis Functions neural network. The above cell-phone source recognition directly processes the original recording file. Since the silent segment contains the same device information as the original speech files, and is not affected by factors such as speaker emotion, voice, intonation and speech content, some researchers began to extract features from the silent segment to characterize the recording device. C. Hanilci et al. [4] extracted MFCC and LFCC features from the silent segment. The results showed that the MFCC features have the highest recognition rate under SVM, and the recognition rates were 98.39% and 97.03%, respectively, on the two databases.

In addition to the cepstral coefficients, power-normalized cepstral coefficient (PNCC) gradually entered the field of source cell-phone identification. Zou et al. [5] used the Universal Background Model of Gaussian Mixture Model (GMM-UBM) classifier to compare MFCC and PNCC in terms of source cell-phone recognition performance. Experiments showed that MFCC is more effective than PNCC. The recognition rates of the two databases reached 92.86% and 97.71%, respectively. Wang et al. [6] extracted an improved PNCC feature from the silent segment, which uses long-term frame analysis to remove the influence of background noise. GMM-UBM was set as the baseline system, which was improved by two-step discriminative training. The experimental results indicated that the average accuracy for 15 kinds of devices was 96.65%.

Although these features have also achieved good results in the field of source cell-phone identification, most of these cepstral coefficients are constructed based on the perception characteristics of the human ear. Researchers hope to find features that can characterize the inherent characteristics of the device and use them as a fingerprint for the device. Some scholars have begun to extract features directly from the spectrum of the Fourier transform domain as distinguishing features of mobile phones. C. Kotropoulos et al. [7] proposed a new source cell-phone identification algorithm, which uses the sketches of spectral features (SSFs) as an intrinsic fingerprint. By applying a sparse-representation-based classifier to the SSFs, identification accuracy exceeded 95% on a set of 8 telephone handsets from the Lincoln-Labs Handset Database. Jin et al. [8] proposed a method for extracting the noise of the recording device from the silent segment. The spectral shape features and spectral distribution features were extracted from the device noise. The features obtained by combining the two features were the best, and recognition rates reached 89.23% and 94.53%, respectively, for the two databases. Qi et al. [9] obtained the noise signal by de-noising using the spectral subtraction method and used the Fourier histogram coefficient of the noise signal as the input for the deep model classifier. In comparing the recognition effects of three different deep learning classifiers—SOFTMAX,

Multilayer perceptron (MLP) and CNN—CNN performed well, and the voting model combined with multiple classifiers had the best effect, with a recognition rate reaching 99%. Recently, Luo et al. [10] proposed a new feature—the band energy difference feature—which is obtained by processing the difference between the energy values of the Fourier transform of the speech file. This feature not only has low computational complexity, but it is also highly distinct for different mobile devices. It reached an accuracy of over 96% using SVM.

Although most source cell-phone identification systems have good accuracy, they have certain limitations. The objects they identify are almost clean speech files (nearly no environmental noise). Few studies have considered noise attacks. In actual life, the speech files that need to be identified are usually recorded in a variety of different noisy environments, and the environmental noise affects the accuracy of recognition. Therefore, the identification of the source cell-phone in a noisy environment is more realistic and challenging. Based on this, this paper proposes a source cell-phone identification algorithm suitable for noisy environments. This algorithm uses the spectrum distribution feature of the constant Q transform domain as the device fingerprint, and uses the multi-scene training method to train the CNN model for source cell-phone identification.

The rest of paper is set out as follows: Section 2 analyzes the differences of speech files recorded by different brands of cell-phone and different models of cell-phone from the same brand; Section 3 presents the spectrum distribution features of the CQT domain proposed in this paper by device difference analysis and two traditional features—MFCC and LFCC; four kinds of classifiers and a cell-phone source identification algorithm flow chart are introduced in Section 4; Section 5 describes the construction process of the basic speech databases and the noisy speech databases; and Section 6 gives the experimental results. Lastly, we conclude this paper in Section 7.

2. Device Difference Analysis

A spectrogram is a visual representation of the spectrum of a speech signal which changes with time. In order to study the differences between speech files recorded by different cell-phones in different frequency bands, Figure 1 shows short-term Fourier transform (STFT) domain spectrograms of speech files recorded by the same speaker simultaneously with 8 cell-phones in a quiet office environment. As can be seen from the figure, the spectrograms of speech files recorded by different brands of cell-phones vary greatly. For example, HuaweiMate7's energy is rapidly reduced near 0.7 kHz, but the decrease of Mi4 is near 1 kHz. The energy distribution of other bands' cell-phones and the frequency band of sudden changes in energy are also different. However, for different models of the brand Apple, the spectrograms are very similar.

To analyze the frequency difference of different models of cell-phones from the same brand, Figure 2 plots the spectrograms of speech files recorded by different models of Apple cell-phones. Although the four images are very similar, with a rapid energy change at around 1.5 kHz, there are still some differences. For example, the iPhone 6 has lower energy in the 0–1.5 kHz band than the other three phones. The iPhone 5 and iPhone 6s have distinct peaks in energy at around 1.2 kHz and 1 kHz, respectively, and the other two phones do not.

According to the above, different brands of cell-phones have large differences and are easy to distinguish. Therefore, the key to source cell-phones identification lies in the identification of different models of cell-phones of the same brand. That is to say, the identification of cell-phones depends on whether it is possible to distinguish well between differences in the middle and low frequencies in the recording equipment.

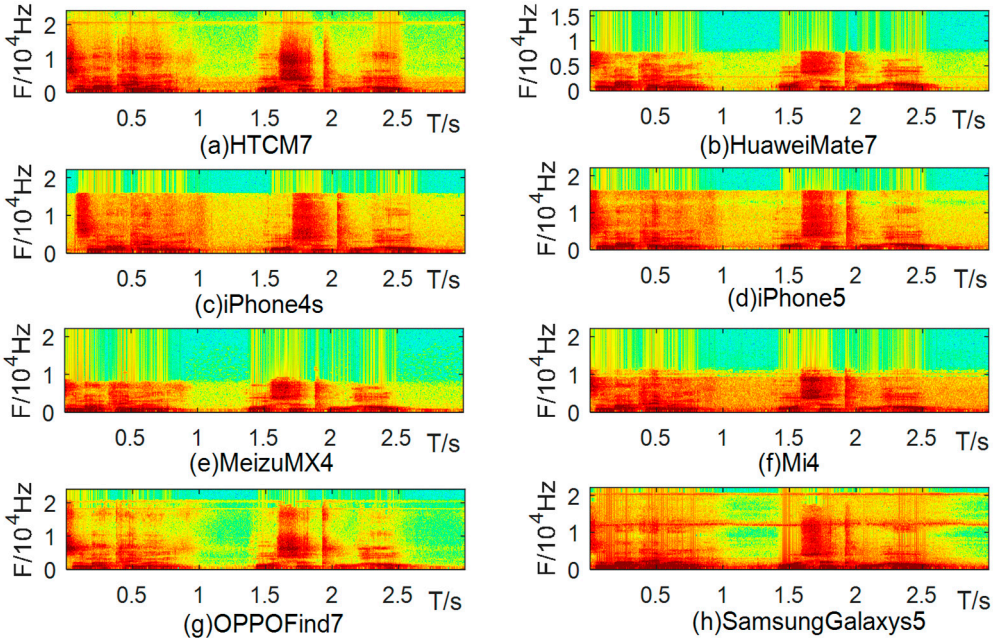

Figure 1. Spectrogram of speech files recorded by different brands of cell-phones.

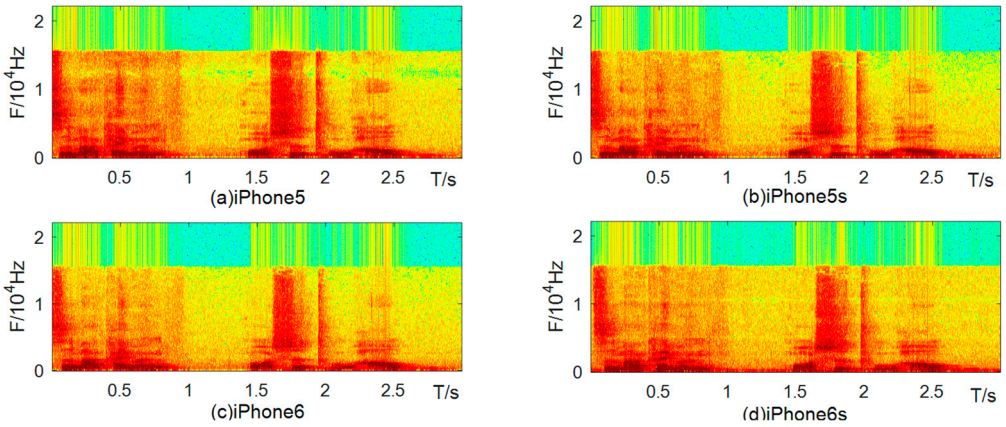

Figure 2. Spectrograms of speech files recorded by different models of Apple cell-phones.

3. Feature Extraction

3.1. Spectral Distribution Features of the CQT Domain

Based on the analysis of the difference in the spectrogram from the Fourier transform (STFT) domain of the speech files recorded by different cell-phones in Section 2, the difference between different brands is obvious and the similarity between different models of the same brand is high, with subtle differences only in the middle- and low-frequency bands. This paper chooses to construct features from the constant Q transform (CQT) domain that can effectively distinguish different recording devices. Compared with STFT with a fixed time-frequency resolution, CQT has a higher frequency resolution at low frequencies and a higher time resolution at high frequencies. To capture these variations in the spectrogram in the CQT domain, this paper selects the spectral distribution features to describe the characteristics of the spectrum.

The calculation process of the spectral distribution features of the CQT domain is given below.

(1) If the time domain signal of a speech files is $x(n)$, and the frequency domain signal after the CQT is $X^{CQT}(k)$, $X^{CQT}(k)$ is defined by:

$$X^{CQT}(k) = \sum_{n=1}^{N_k} x(n) w_{N_k}(n) e^{-j2\pi \frac{f_k}{f_s} n} \quad (1)$$

where k = 1, 2, ..., K is the frequency bin index; f_s is the sampling rate; f_k is the center frequency of bin k, which is exponentially distributed and is defined as

$$f_k = f_1 \times 2^{\frac{k-1}{B}} \quad (2)$$

where B is the number of bins per octave, f_1 is the center frequency of the lowest frequency bin and is computed according to:

$$f_1 = \frac{f_{max}}{2^{(K-1)/B}} \quad (3)$$

$$f_{max} = \frac{f_s}{2} \quad (4)$$

$w_{N_k}(n)$ is a window function (Hanning window); from high frequency to low frequency, with the increasing frequency resolution, the time resolution will gradually be sacrificed. Therefore, the window length N_k varies with k and is inversely proportional to k, namely:

$$N_k = \frac{f_s}{f_k} \times Q \quad (5)$$

The Q-factor is a constant independent of k and is defined as the ratio of the center frequency to the bandwidth: $Q = \frac{f_k}{BW_k} = \frac{f_k}{f_{k+1} - f_k} = \left(2^{\frac{1}{B}} - 1\right)^{-1}$.

(2) For the frequency value $X_i(k)$ of the i-th frame at the kth frequency point, the amplitude $Y_i(k)$ of $X_i(k)$ is computed as follows:

$$Y_i(k) = X_i(k) \times \overline{X_I(k)} \quad (6)$$

(3) Spectral distribution features:

$$SSF(k) = \frac{1}{T_k} \sum_{i=1}^{T_k} \mathrm{In} Y_i(k) \quad (7)$$

where T_k represents the total number of frames of the speech in the kth frequency band, k = 1, 2, ... , K. Therefore, for a speech file, its spectral distribution features comprise a $1 \times K$ vector. K is set as 420 in this paper.

3.2. Traditional Features (MFCC, LFCC)

Almost all of the features in cell-phone source recognition are extracted from the Fourier transform domain of the speech signal. The most popular features for recognition systems are the MFCC and LFCC. Their extraction procedure is shown in Figure 3. Firstly, the speech signal is divided into overlapping frames, and each frame is windowed using an appropriate window function. Then, the power spectrum is computed using the fast Fourier transform (FFT), which is then smoothed with a bank of triangular filters whose center frequencies are uniformly spaced on the different scales. Finally, logarithmic filter bank outputs are converted into features by taking the discrete cosine transform (DCT). In this paper, we use 30 millisecond frames with a 15 millisecond overlap and a Hamming window. For every speech file, the dimensions of MFCC and LFCC are both set to 24.

The difference between MFCC and LFCC lies in different frequency scale to locate triangular filters. MFCC uses a Mel-frequency scale, but the triangular filters are linearly spaced for LFCC.

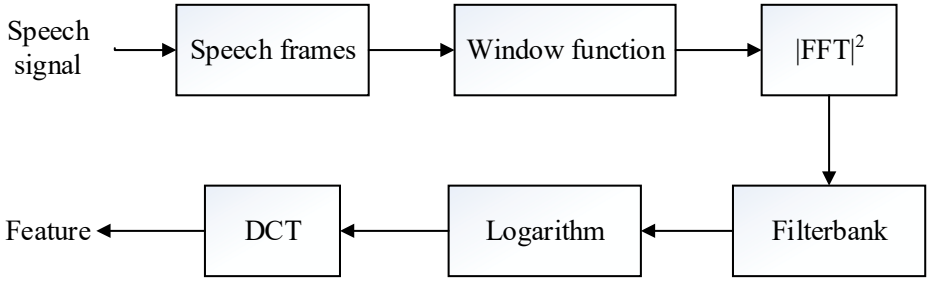

Figure 3. MFCC, LFCC extraction process.

4. Classifiers and Algorithm Introduction

4.1. SVM

The Support Vector Machine (SVM) is a supervised learning classification algorithm for solving the two-class problem. Its basic model is to find the best-separated hyperplane in the feature space, so that the positive and negative sample intervals on the training set are as large as possible. SVM can be used to solve linear problems, and it can also be used to solve nonlinear problems after introducing a kernel method. For the n-classification problem in this paper, $n \times (n-1)/2$ SVM classifiers are integrated for classification, and the kernel function selects the Gaussian kernel.

4.2. RF

When Random Forest (RF) is used as a classifier, it is an integrated classification algorithm that relies on the voting choices of multiple decision trees to determine the final classification result. RF has two characteristics, sample randomness and feature randomness, ensuring that there is no over-fitting phenomenon when classifying Sample randomness means that if the training set size is x, for each decision tree, training samples (also for x) are randomly and regressively extracted from the training set using the bootstrapping method. Feature randomness means that when training the nodes of each decision tree, the features used are randomly selected from all the features according to a certain proportion. By calculating the amount of information contained in each feature, the feature with the highest classification ability is selected for node splitting. In this paper, the CART algorithm is used to generate the decision tree, and the Gini index is used as the criterion for selecting the optimal feature and the splitting point.

4.3. CNN

Convolutional Neural Network (CNN) is a multi-layer neural network consisting mainly of a convolutional layer, a pooling layer, a nonlinear activation layer, and a fully connected layer. The nonlinear modeling capability of CNN makes it an excellent classifier. The CNN network structure adopted in this paper is shown in Figure 4. As you can see, it uses dropout, which is a regularization method to prevent overfitting. The basic idea of dropout is to randomly drop out some neurons during the training of the deep learning network. The model can be made more robust because it does not rely too much on local features (because local features are likely to be discarded). This paper uses the method of random gradient descent when training the CNN model.

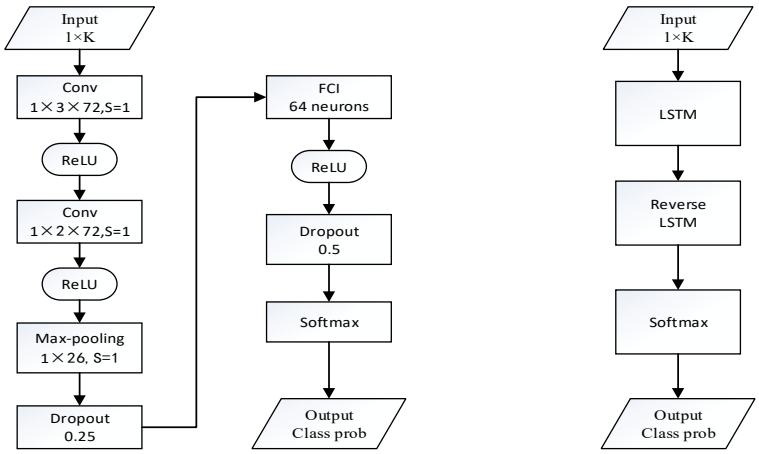

Figure 4. Network framework for different depth learning classifiers. Left: CNN Right: RNN-BLSTM.

4.4. RNN-BLSTM

Recurrent Neuron Network (RNN) is a neural network that models sequence data. The network memorizes previous information and applies it to the calculation of the current output. That is, the nodes between the hidden layers are no longer connectionless but connected. In addition, the input of the hidden layer includes not only the output of the input layer but also the output of the hidden layer at the previous moment. Long Short-Term Memory Neural Network (LSTM) is a special type of RNN that is more advantageous than basic RNN in sequence generation and sequence tagging. Each node in the RNN–LSTM hidden layer is represented by an LSTM structure compared to a normal RNN model with a basic node. The one-way RNN–LSTM only models the forward sequence, which means that the latter sequence cannot affect the modeling of the previous sequence. If the entire sequence in the positive and negative directions can be used at the same time, the accuracy of the model should be improved. Therefore, RNN-BLSTM is selected to classify different recording devices. The specific network framework used is shown in Figure 4.

4.5. Multi-Scene Training Recognition Systems

In this paper, the multi-scene training method is used to enhance the noise robustness of the source cell-phone identification system. The specific flow chart of this system is shown in Figure 5. The traditional single-scenario training method only uses clean speech files to extract the distinguishing features of the device, and then uses those features to establish a recognition model. When the multi-scene training method is used to build the model, the training set not only has clean speech files but also noisy speech files containing different noise types and different noise intensities. This model can learn the effect of noise on the differences in speech recorded by different recording devices, making the model more robust.

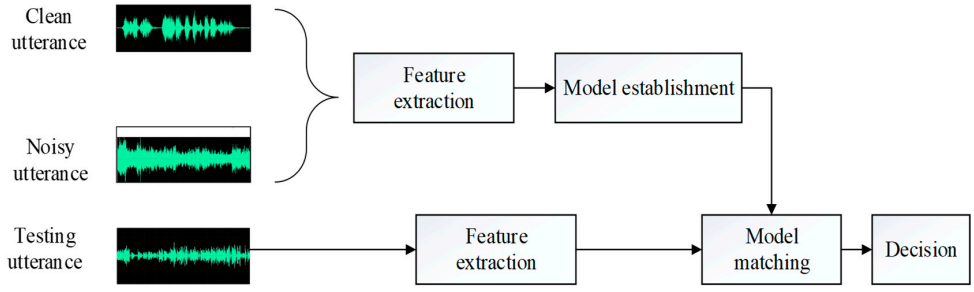

Figure 5. Source cell-phone identification algorithm block diagram of multi-scene training.

5. Databases Construction

5.1. Basic Speech Databases

In the experiment, we used 24 cell-phones from 7 brands to research source cell-phone recognition, and their specific information is as shown in Table 1. To make the experimental results comparable, we used two different basic speech databases, the TIMIT Recaptured Database (TIMIT-RD) and the CKC speech database (CKC-SD) [11], to investigate the performance of our cell-phone recognition system under different conditions. They were built using the cell-phone devices of Table 1.

TIMIT-RD is a speech database comprising speech files ripped from the TIMIT [12] database, and these speech files include 1600 speech samples of 160 people (half each for men and women) which were played back through a high-fidelity speaker (PhilipsDTM3500, manufacturer: Philips Investment Co., Ltd., Zhongshan, China) in a quiet office environment. For each cell-phone, we used 800 utterances for training and the remaining 800 utterances for testing. The CKC-SD database is a second database constructed by recording speech spoken by 12 speakers (half of which were male) in a quiet office environment. All cell-phones were placed in a circular arc around the speaker for simultaneous recording. Each speaker recorded two speech segments with a duration of more than 5 min, based on a normal speech rate and intonation. One speech segment was fixed content, and each person's recorded content was the same. The other section used the form of question and answer, and the content recorded by each person was different. Half of the recording (5 min), which was segmented into 3 s long chunks, was used to train each phone, and the remaining 5 min portion was segmented into 3 s long chunks for testing.

Table 1. Cell-phone list.

Class ID	Brand	Model	Class ID	Brand	Model
H1	HTC	D610t	A1	iPhone	iPhone 4s
H2		D820t	A2		iPhone 5
H3		One M7	A3		iPhone 5s
W1	Huawei	Honor6	A4		iPhone 6
W2		Honor7	A5		iPhone 6s
W3		Mate7	Z1	Meizu	Meilan Note
O1	OPPO	Find7	Z2		MX2
O2		Oneplus1	Z3		MX4
O3		R831S	M1	Mi	Mi 3
S1	Samsung	Galaxy Note2	M2		Mi 4
S2		Galaxy S5	M3		Redmi Note1
S3		Galaxy GT-I8558	M4		Redmi Note2

5.2. Noisy Speech Databases

To study the robustness of the source cell-phone identification system in noisy environments, different types of noises at a variety of signal-to-noise ratios need to be added to the two basic speech databases to simulate the actual noise scene. When adding noise to the underlying databases, we used the filtering and noise addition tool (FaNT, version), which is an open-source tool that follows ITU's noise addition and filtering. The noise signal was selected from five noise types—white noise, babble noise, street noise, cafe noise and Volvo noise—in the NOISEX-92 noise database, and for each type of noise, three signal-to-noise ratio (SNR) levels, i.e., 0 dB, 10 dB and 20 dB were considered. Therefore, each basic speech database constituted 15 noisy databases with different noise intensities and different types of noise. The reasons for choosing these five types of noise are as follows: (1) the energy of white noise is evenly distributed over the frequency components. Although it rarely represents the actual situation, it is a commonly used noise when studying robust speech processing methods; (2) Babble noise is one of the most difficult types of noise in speaker applications with multiple

speakers, which occurs every day in any crowded place; (3) Streets, cafe and Volvo noises are other types of noise that often occur in our daily lives.

6. Experiments

6.1. Experimental Setup

For the multi-scene training recognition system used in this paper, the types of speech files included in the training set and the test sets are as follows:

Train: clean, white (0,10,20 dB), babble (0,10,20 dB), street (0,10,20 dB).

Test: clean, white (0,10,20 dB), babble (0,10,20 dB), street (0,10,20 dB), cafe (0,10,20 dB), Volvo (0,10,20 dB).

The training set used in the training phase includes not only clean speech files in the basic speech database, but also three noisy speech files with three signal-to-noise ratio (SNR) levels of white, babble, and street noise types in the noisy speech database. When testing the model trained in the training set, 16 test sets are used, each of which includes one type of speech file, 10 of which comprise the clean speech file and 9 different noisy speech files with different noise types and noise intensities that were in the training set (seen noisy speech); the remaining 6 test sets are noisy speech files for the two noise types that were not in the training set (unseen noisy speech). This is advantageous for detecting whether the model established by the multi-scene training is universal; that is, whether it can conduct effective source cell-phone recognition on the speech files of the unseen noisy scenarios.

6.2. Parameter Setup

When the CQT converts the time domain signal of the speech file into the frequency domain, K determines the accuracy of the frequency domain information of the speech signal, so the magnitude of K has a certain influence on the recognition performance. As shown by Equation (3):

$$K = \left\lfloor B \times \log_2 \frac{f_{max}}{f_1} \right\rfloor \tag{8}$$

$$OCT = \left\lfloor \log_2 \frac{f_{max}}{f_1} \right\rfloor \tag{9}$$

where $\lfloor x \rfloor$ represents the largest integer less than or equal to x, OCT is the number of octaves (generally an integer less than or equal to 9), and B is the number of frequency points per octave (generally about 100). So the size of K is determined by B and OCT.

Figure 6 shows that the performance of the spectral distribution features of different OCT values when B is 100. The accuracy of the device recognition experiments are compared by using two traditional classifiers—SVM and RF, respectively—on the CKC-SD database. It can be seen from the figure that different OCT values have similar effects on the recognition of different test sets under the two traditional classifiers. As the OCT value increases, the accuracy of the noisy speech files containing white noise and babble noise gradually increases, but that of the speech files containing Volvo noise drops sharply, and the remaining speech types are not significantly affected, except for cafe noise. Considering the complexity of the algorithm and the influence of different OCT values on the accuracy of different noise types, the OCT value is set to 7.

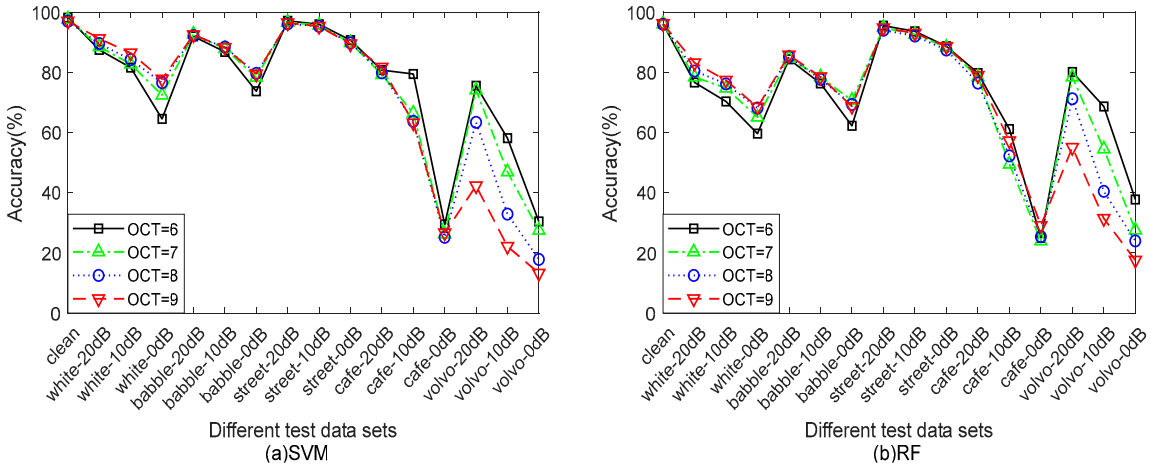

Figure 6. Comparison of different OCT values.

Figure 7 shows the comparison of different B values for the influence of device recognition when the OCT value is 7. Experiments were carried out using two traditional classifiers—SVM and RF—on the CKC-SD database. From the figure, with an increase in B value, there is little effect on the recognition of clean speech files and noisy speech files containing seen noise types, and the recognition rate of speech files containing unseen noise types is only slightly improved. In short, the value of B has little effect on the accuracy of source identification, so we choose 60 as the value of B. Therefore, the K value is 420 (7×60) in the paper. For a speech file, number of its spectral distribution features is 420.

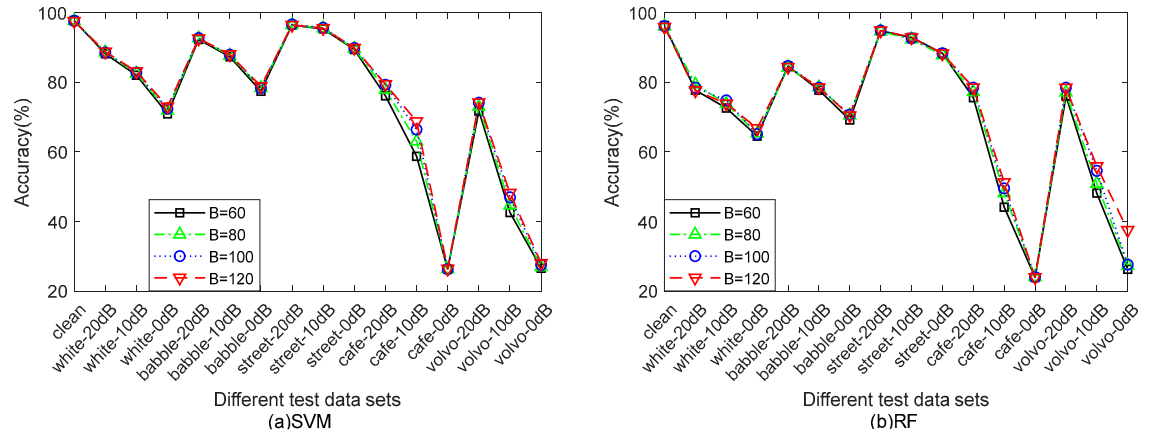

Figure 7. Comparison of different B values.

6.3. Comparison of Features

To compare the influence of the proposed features and traditional features on the performance of source cell-phone recognition, Figure 8 plots the accuracy of four features—MFCC, LFCC, SSF (STFT) and SSF (CQT)—on different test sets under SVM. As can be seen from the figure, these four features have a good recognition effect on clean speech files, but with the addition of noise, the accuracy decreases, and the performance gets worse with increasing noise intensity. Secondly, for the same noise intensity, the recognition of seen noisy speech files is significantly better than the unseen noisy speech files. The recognition rate of the traditional features—MFCC and LFCC—for noisy speech files decreases sharply with the increase in noise intensity; the situation is even worse for the unseen noisy scenarios, where the highest accuracy is only 80%, and the lowest accuracy is 24%. Therefore, the traditional features are poorly robust to noise. The performance of SSF (STFT) features is generally

worse than the traditional features, but it is superior to traditional features in the case of strong noise intensity. The SSF (CQT) feature is more robust than the other features. It is obviously better than MFCC, LFCC, and SSF (STFT) for clean speech and seen noisy speech files, with an accuracy higher than 70%. However, the recognition effect of the unseen noisy speech files does not change significantly compared with the other features. For weak noise intensity, the accuracy is slightly improved, while for high noise intensity, the accuracy is reduced.

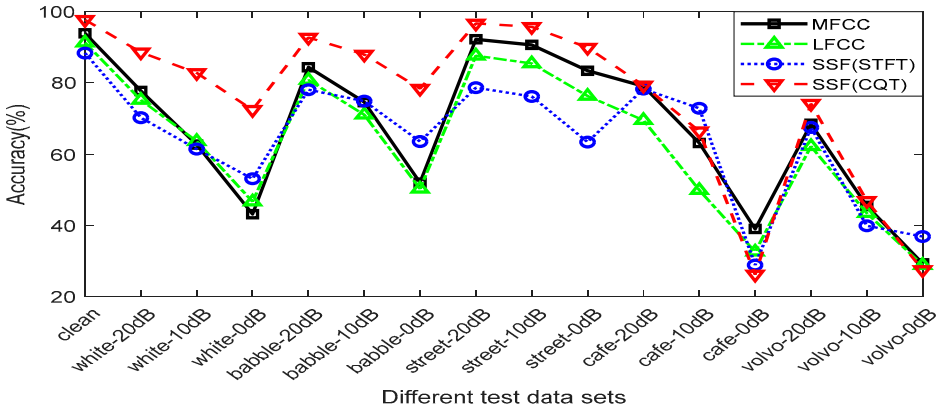

Figure 8. Comparison of accuracy of different features under SVM.

In general, the SSF (CQT) feature is significantly superior to the other features as the device fingerprint. The MFCC, LFCC and SSF (STFT) features are extracted from the STFT domain, while SSF (CQT) is derived from the CQT domain. Therefore, the frequency domain information obtained by using different time-frequency transform methods of speech signals is different, leading to a difference in accuracy. CQT is more suitable for source cell-phone recognition than the STFT.

Tables 2 and 3, respectively, show the specific classification results of the MFCC and SSF (CQT) features on the clean test set from CKC-SD. In the tables, AL is the actual device model in which the speech files are recorded, and PL indicates the predicted device model. It can be seen from Table 2 that the average accuracy of MFCC for the 24 devices is 92%. The overall performance of MFCC is good, but the accuracy varies greatly for different device models. The recognition rate of Meizu and Xiaomi is almost 100%. The recognition rate is the lowest for two models (D610t, D820t) of HTC, at 56% and 79%, respectively. Like Huawei and Apple, three models of HTC are also misjudged within the brand. The misclassification of Xiaomi and Samsung is mainly misjudged within the brand, but also includes a small number of misjudgments outside the brand. It can be seen from Table 3 that the average accuracy of SSF (CQT) for 24 devices is 98%, which is 6 percentage points higher than that of MFCC. This feature is almost perfect for the recognition of Meizu, Xiaomi, OPPO and Samsung brands. The wrong scores for HTC, Huawei, and Apple are misjudgments within the brand, and the accuracy is improved compared to MFCC.

Table 4 shows the classification results of MFCC and SSF (CQT) features for different brands on different test sets. Regardless of whether examining the clean speech test set or the noisy speech test set containing white noise, the SSF (CQT) feature improves the accuracy for each brand compared to the MFCC. This confirms that the high resolution of the low-middle frequency band in CQT can improve the recognition performance of different models of cell-phones of the same brand.

The above experimental results and analysis indicate that: SSF (CQT) can be used to determine the unique identity information of a specific model of device, and can effectively identify the recording equipment for both clean speech files and noisy speech files.

Table 2. Specific classification results of MFCC for the clean test set (%).

	PL (Predict Label)																						
AL	**H1**	**H2**	**H3**	**W1**	**W2**	**W3**	**A1**	**A2**	**A3**	**A4**	**A5**	**Z1**	**Z2**	**Z3**	**M1**	**M2**	**M3**	**M4**	**O1**	**O2**	**O3**	**S1**	**S2**	**S3**
H1	0.56	0.44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H2	0.21	0.79	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H3	0.04	0.05	0.91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W1	0	0	0	0.97	0	0.03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W2	0	0	0	0.05	0.83	0.12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W3	0	0	0	0.10	0	0.89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0
A1	0	0	0	0	0	0	0.88	0	0.04	0.08	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A2	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A3	0	0	0	0	0	0	0.03	0	0.85	0.12	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A4	0	0	0	0	0	0	0.33	0	0.11	0.56	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A5	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
Z1	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0
Z2	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0
Z3	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0
M1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0	0	0	0
M2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0	0	0
M3	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0.99	0	0	0	0	0	0	0
M4	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0	0	0.83	0.15	0	0	0.01	0	0
O1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0	0	0.99	0	0	0	0	0
O2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0
O3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0
S1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.04	0	0	0	0.01	0	0	0.95	0	0
S2	0	0	0	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.98	0
S3	0	0	0	0	0	0	0	0	0	0	0	0.08	0	0.02	0	0	0	0	0	0	0	0	0	0.90

Table 3. Specific classification results of SSF (CQT) for the clean test set (%).

	PL (Predict Label)																						
AL	**H1**	**H2**	**H3**	**W1**	**W2**	**W3**	**A1**	**A2**	**A3**	**A4**	**A5**	**Z1**	**Z2**	**Z3**	**M1**	**M2**	**M3**	**M4**	**O1**	**O2**	**O3**	**S1**	**S2**	**S3**
H1	0.87	0.13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H2	0.14	0.86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H3	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W1	0	0	0	0.97	0.02	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W2	0	0	0	0.01	0.96	0.03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W3	0	0	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A1	0	0	0	0	0	0	0.98	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A2	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A3	0	0	0	0	0	0	0.05	0	0.92	0.02	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0
A4	0	0	0	0	0	0	0.04	0.04	0	0.92	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A5	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
Z1	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0
Z2	0	0	0	0	0.01	0	0	0	0	0	0	0	0.99	0	0	0	0	0	0	0	0	0	0	0
Z3	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0
M1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0	0	0	0
M2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0	0	0
M3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0	0
M4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0	0	0
O1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0	0
O2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0	0
O3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0	0
S1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	0
S2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0
S3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00

Table 4. Classification results of MFCC and SSF (CQT) for different brands.

Brands	Clean		White_0dB	
	MFCC	SSF (CQT)	MFCC	SSF (CQT)
HTC	75.33%	91.00%	49.17%	70.33%
Huawei	89.67%	97.67%	68.83%	77.17%
iPhone	85.80%	96.40%	36.10%	56.70%
Meizu	100%	99.67%	33.31%	69.50%
Mi	95.50%	100%	31.13%	85.25%
OPPO	99.67%	100%	31.00%	82.83%
Samsung	94.34%	100%	43.33%	71.67%

6.4. Comparison of Classifiers

Figure 9 compares the performance of SSF (CQT) under four different classifiers. It can be seen from the figure that the traditional SVM and RF classifiers have an almost identical recognition effect for clean speech files, but there are differences for noisy speech files. The classification effect of RNN on the clean speech test set and the noisy speech test sets with white noise, babble noise, and street noise is significantly worse than that of the traditional classifiers, but the recognition effect in unseen noisy speech is better than the traditional classifiers, especially for Volvo noise, where the highest accuracy shows an increase of about 20%. Surprisingly, the accuracy of CNN on the 16 test sets is higher than that of the other three classifiers, especially for the speech test set of unseen noisy scenarios, the performance of which is greatly improved. In the test sets of cafe and Volvo noise with different noise intensities, most of the accuracies are higher than 90%, with the lowest accuracy also being greater than 70%.

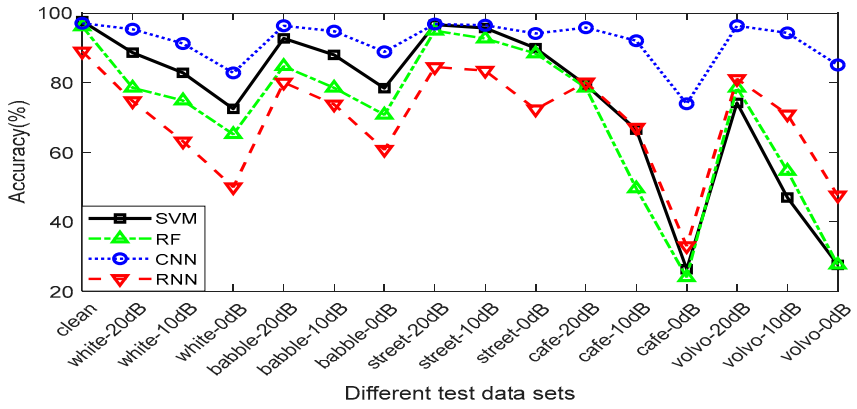

Figure 9. Comparison of accuracy of different classifiers using SSF (CQT).

Therefore, the performance of the deep-learning CNN classifier is very prominent, not only maintaining a good performance on clean speech files, but also having a good recognition effect on 15 kinds of noisy speech test sets. Even if the training concentration does not include these speech files with cafe noise and Volvo noise, CNN can also distinguish recording devices from noisy speech files containing these two kinds of noise, and can achieve accuracy comparable to that for seen noisy speech. Therefore, CNN is more suitable for source identification in noisy environments.

6.5. Comparison of Single-Scene and Multi-Scene Training

To verify the effectiveness of the multi-scene training method, Table 5 compares the performance of single-scene and multi-scene training methods on the two databases CKC-SD and TIMIT-RD, respectively. The features and classifiers use SSF (CQT) and CNN, respectively. It can be seen from the figure that when the testing speech consists of clean speech files, the recognition rate of the

multi-scene training algorithm is higher than that for the single-scene algorithm in the two databases, indicating that if noisy speech files are added to the training set, the recognition effect will be improved. Secondly, when the testing speech files consist of noisy speech, the accuracy of the multi-scene training recognition algorithm is greatly improved for the two speech databases compared to the single-scene training method, especially for high-intensity noisy speech, the accuracy of which can be increased by up to 60%.

Table 5. Comparison of accuracy of single-scene and multi-scene training.

Test Data Sets	Single		Multiple	
	CKC-SD	TIMIT-RD	CKC-SD	TIMIT-RD
Seen Noisy Scenarios				
clean	95.47%	98.89%	97.08%	99.29%
white_20dB	54.80%	58.80%	95.35%	96.31%
white_10dB	36.11%	35.11%	91.25%	91.99%
white _0dB	18.50%	16.50%	82.79%	84.57%
babble_20dB	76.71%	77.71%	96.35%	97.54%
babble 10dB	49.92%	50.92%	94.79%	96.03%
babble_0dB	26.77%	29.77%	88.85%	90.23%
street_20dB	97.86%	98.86%	96.85%	98.44%
street_10dB	86.27%	87.27%	96.50%	97.40%
street_0dB	54.81%	52.81%	94.13%	93.47%
Unseen Noisy Scenarios				
cafe_20dB	88.33%	89.99%	95.81%	96.56%
cafe_10dB	56.25%	61.25%	92.04%	94.63%
cafe_0dB	30.75%	27.75%	73.90%	76.43%
volvo_20dB	92.33%	93.33%	96.35%	96.34%
volvo_10dB	71.98%	76.98%	94.30%	92.21%
volvo_0dB	45.75%	46.75%	85.06%	88.06%

The multi-scene training recognition algorithm using the features proposed in this paper and the CNN classifier not only achieves a good performance in the seen noise-scene speech files, but also has considerable accuracy in the unseen noise-scene speech files. Therefore, training the model using the multi-scene training method can solve the carrier mismatch problem of the single-scene training method.

6.6. Comparison of Different Identification Algorithms

To comprehensively evaluate the source cell-phone identification algorithm proposed in this paper, we compare the recognition algorithm from the Reference [10] and this paper's recognition algorithm using multi-scene training method and testing on our speech databases. The number of training and test speech files is the same as that used in this article. Recognition algorithm of Reference [10] extracted the sub-band energy difference feature as a distinguishing feature, which is obtained by performing differential processing on the power value of the original speech file after Fourier transform and uses SVM as a classifier. Their parameter settings are consistent with Reference [10].

Table 6 is a comparison of the accuracy of the source cell-phone identification of the algorithm presented in Reference [10] and the algorithm proposed in this paper. It can be seen from the table that the two algorithms are almost equivalent in performance for clean speech files, but for noisy speech files, this paper's algorithm is superior to the algorithm presented in Reference [10], especially for noisy speech of unseen noise type.

Table 6. Comparison of accuracy of different algorithms

Test Data Sets	This Paper		Reference [10]	
	CKC-SD	TIMIT-RD	CKC-SD	TIMIT-RD
Seen Noisy Scenarios				
clean	97.08%	99.29%	97.04%	98.69%
white_20dB	95.35%	96.31%	89.60%	88.60%
white_10dB	91.25%	91.99%	84.29%	82.82%
white _0dB	82.79%	84.57%	76.62%	72.46%
babble_20dB	96.35%	97.54%	92.29%	92.73%
babble 10dB	94.79%	96.03%	88.46%	87.98%
babble_0dB	88.85%	90.23%	79.73%	78.40%
street_20dB	96.85%	98.44%	96.19%	96.69%
street_10dB	96.50%	97.40%	95.38%	95.73%
street_0dB	94.13%	93.47%	89.81%	89.90%
Unseen Noisy Scenarios				
cafe_20dB	95.81%	96.56%	79.74%	79.31%
cafe_10dB	92.04%	94.63%	63.73%	66.46%
cafe_0dB	73.90%	76.43%	25.19%	26.27%
volvo_20dB	96.35%	96.34%	64.32%	74.17%
volvo_10dB	94.30%	92.21%	32,92%	46.94%
volvo_0dB	85.06%	88.06%	17.85%	27.54%

7. Conclusions

Currently, all source cell-phone identification algorithms use cepstral coefficients, features based on cepstral coefficients, or extract features directly from the spectrum of the Fourier transform domain as the distinguishing feature of the mobile phone. They have good performance, but the recognition objects of source cell-phone recognition are almost always speech files recorded in a quiet environment (can be considered as no scene noise). When the speech file contains scene noise, the performance of the source cell-phone recognition algorithm drops sharply, and as the noise intensity increases, the recognition becomes worse and worse, so the noise robustness of these algorithms is poor. Considering that speech files that need to be recognized usually contain scene noise, and that the traditional recognition algorithm has poor noise robustness, this paper proposes a source cell-phone recognition algorithm suitable for noisy environments that has very good recognition performance for clean speech files and noisy speech files. It has strong noise robustness.

Through analysis, this paper finds that the difference between different brands of cell-phones is mainly at high frequency, and the difference is obvious, and is easy to distinguish. However, cell-phones of different models of the same brand only have slight differences in the middle and low frequencies, which are difficult to distinguish. Therefore, the algorithm extracts the spectrum distribution feature of the CQT domain of the speech file as the device fingerprint. When the CQT converts the speech time domain signal to the frequency domain, the frequency resolution increases from high frequency to low frequency, such that the low-frequency information amplifies the slight differences between different models of cell-phones of the same brand, and enhances the recognition of different models of mobile phones of the same brand. The features of the traditional recognition algorithm are extracted from the Fourier transform domain. Fourier transformation uses a fixed frequency resolution, and the representation of the low-frequency information of speech is less accurate than CQT. Secondly, the multi-scene training method adopted in this paper can not only improve the accuracy of clean speech files, but can also improve the performance for noisy speech files with seen noise types. Despite this, the recognition performance for noisy speech files with unseen noise types is not improved. Finally, this paper uses CNN (deep learning classifier) as a classifier. Compared with the machine learning classifier used in the traditional source recognition algorithm, it not only improves the recognition of clean speech files and noisy speech files with seen noise types, but also greatly

improves the recognition rate of noisy speech files with unseen noise types. The most important point is that the recognition effect of noisy speech files with unseen noise types is equivalent to noisy speech files of seen noise types. Therefore, the algorithm proposed in this paper is much more robust than traditional source cell-phone recognition algorithms.

Nevertheless, in high-intensity noise environments, the proposed algorithm cannot distinguish the equipment well, and does not consider the distinction between individual recording devices of the same brand and model.

Author Contributions: Conceptualization, W.R. and Y.D.; Methodology, Q.T.; Validation, L.L.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. U1736215, 61672302), Zhejiang Natural Science Foundation (Grant No. LZ15F020002, LY17F020010), Ningbo Natural Science Foundation (Grant No. 2017A610123), Ningbo University Fund (Grant No. XKXL1509, XKXL1503), Mobile Network Application Technology Key Laboratory of Zhejiang Province (Grant No. F2018001).

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

References

1. Hanilci, C.; Ertas, F.; Ertas, T. Recognition of Brand and Models of Cell-Phones from Recorded Speech Signals. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 625–634. [CrossRef]
2. Hanilçi, C.; Ertas, F. Optimizing Acoustic Features for Source Cell-Phone Recognition Using Speech Signals. In Proceedings of the First ACM Workshop on Information Hiding and Multimedia Security, Montpellier, France, 17–19 June 2013; pp. 141–148.
3. Kotropoulos, C.; Samaras, S. Mobile Phone Identification Using Recorded Speech Signals. In Proceedings of the 19th International Conference on Digital Signal Processing, Hong Kong, China, 20–23 August 2014; pp. 586–591.
4. Hanilçi, C.; Kinnunen, T. Source Cell-Phone Recognition from Recorded Speech Using Non-speech Segments. *Digital Signal Process.* **2014**, *35*, 75–85. [CrossRef]
5. Zou, L.; Yang, J.; Huang, T. Automatic cell phone recognition from speech recordings. In Proceedings of the 2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP), Xi'an, China, 9–13 July 2014; pp. 621–625.
6. He, Q.; Wang, Z.; Rudnicky, A.I.; Li, X. A Recording Device Identification Algorithm Based on Improved PNCC Feature and Two-Step Discriminative Training. *Electron. J.* **2014**, *42*, 191–198.
7. Kotropoulos, C. Telephone Handset Identification Using Sparse Representations of Spectral Feature Sketches. In Proceedings of the 2013 International Workshop on Biometrics and Forensics (IWBF), Lisbon, Portugal, 4–5 April 2013; pp. 1–4.
8. Jin, C.; Wang, R.; Yan, D.; Tao, B.; Chen, Y.; Pei, A. Source Cell-Phone Identification Using Spectral Features of Device Self-noise. In Proceedings of the 15th International Workshop on Digital Watermarking (IWDW), Beijing, China, 17–19 September 2016; pp. 29–45.
9. Qi, S.; Huang, Z.; Li, Y.; Shi, S. Audio Recording Device Identification Based on Deep Learning. In Proceedings of the 2016 IEEE International Conference on Signal and Image Processing (ICSIP), Beijing, China, 13–15 August 2016; pp. 426–431.
10. 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]
11. Jin, C. Research on Passive Forensics for Digital Audio. Ningbo University, 2016; pp. 28–35. Available online: http://cdmd.cnki.com.cn/Article/CDMD-11646-1017871275.htm (accessed on 17 August 2018).
12. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S.; Dahlgren, N.L. *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1*; U.S. Department of Commerce: Gaithersburg, MD, USA, 1993; p. 93.