

## A Novel System for Recognizing Recording Devices from Recorded Speech Signals

Yongqiang Bao<sup>1,\*</sup>, Qi Shao<sup>1</sup>, Xuxu Zhang<sup>1</sup>, Jiahui Jiang<sup>1</sup>, Yue Xie<sup>1</sup>, Tingting Liu<sup>1</sup>  
and Weiye Xu<sup>2</sup>

**Abstract:** The field of digital audio forensics aims to detect threats and fraud in audio signals. Contemporary audio forensic techniques use digital signal processing to detect the authenticity of recorded speech, recognize speakers, and recognize recording devices. User-generated audio recordings from mobile phones are very helpful in a number of forensic applications. This article proposed a novel method for recognizing recording devices based on recorded audio signals. First, a database of the features of various recording devices was constructed using 32 recording devices (20 mobile phones of different brands and 12 kinds of recording pens) in various environments. Second, the audio features of each recording device, such as the Mel-frequency cepstral coefficients (MFCC), were extracted from the audio signals and used as model inputs. Finally, support vector machines (SVM) with fractional Gaussian kernel were used to recognize the recording devices from their audio features. Experiments demonstrated that the proposed method had a 93.4% accuracy in recognizing recording devices.

**Keywords:** Recording device recognition, Mel-frequency cepstral coefficients, support vector machines.

### 1 Introduction

Recognizing recording devices involves extracting features from recorded audio signals and using pattern recognition to recognize which device recorded the audio signal. Mobile phones are the most popular recording devices and they vary by brands and models. Digital audio forensics determines the kind of device used to record audio, and it has been critical in driving the adoption of audio recordings as court-approved evidence.

Recognizing the recording device of an audio signal is the latest development in the field of audio forensics research. The recording module of a recording device is generally composed of an analog front-end, an analog-to-digital (A/D) converter, a noise reduction algorithm, and a compression algorithm. The manufacturers of recording devices

---

<sup>1</sup> Nanjing Institute of Technology, Nanjing, 211167, China.

<sup>2</sup> Department of Informatics, University of Leicester, Leicester, LE1 7RH, UK.

\* Corresponding Author: Yongqiang Bao. Email: jybyq@163.com.

Received: 28 April 2020; Accepted: 16 August 2020.

generally use different analog circuits and digital signal processing algorithms to record audio, which causes audio signals to have different features across different devices.

A vector of audio features contains information relating to voice, content, speaker, recording environment, and recording device. These features overlap in the time and frequency domains, making it difficult for general classification methods to separate them. Research on recognizing recording devices is still in its infancy, and studies have not been conducted on the audio features of various types of mobile phones and recording devices. Furthermore, there have been no effective solutions for constructing a database of the features of recording devices, extracting the audio features of each recording device, and designing a recognition model. A lot of research remains to be done in this field.

The rest of this article is organized as follows. System definitions and the algorithm for recognizing recording devices are described in Section 2. Section 3 details MFCC extraction, and classification methods for recognizing recording devices. The results of experiments are provided in Section 4, and future work and conclusions are discussed in Section 5.

## **2 Related works**

Audio signals contain information, which can be used to recognize the relevant recording device. In 2007, Kreutzer et al. [Kraetzer, Oermann, Dittmann et al. (2007a)] used 14 microphones to record audio signals in 11 rooms, and then used a Bayes classification algorithm to classify the recording devices and recognize recording environments with 75.99% accuracy. In 2014, Aggarwal et al. [Aggarwal, Singh, Roul et al. (2014a)] studied 26 mobile phone models from five manufacturers (including Nokia, Samsung, Blackberry, Sony, and Zen) and established a database of the audio features of these mobile phones. Aggarwal et al. [Aggarwal, Singh, Roul et al. (2014b)] then proposed an audio recognition model that used 24-dimensional MFCC with mixed parameters as input, and trained SVM using sequential minimal optimization. The recognition rate of this model across the five mobile phone brands was 90%, and the average recognition rate for each Nokia model was also 90%. In 2019, the projected Gaussian Supervector (GSV) proposed by Jiang et al. [Jiang and Leung (2019)] achieved a high rate of recognition. If noise in the non-speech segments of a recording is stable and long enough, then power spectrum can be used to recognize the recording device. However, if the non-speech segments of the recording are not long enough or get interrupted by other noises, then power spectrum cannot be used to effectively recognize the recording device.

### **2.1 Non-speech detection**

Audio signals can be divided into speech and non-speech segments. Features of the recording device are contained in both the speech and non-speech segments of an audio signal. The power in the speech signal usually accounts for a large portion of the power in audio signals. In speech signal processing, non-speech detection is also known as endpoint detection.

Endpoint detection generally has five stages, namely, framing, pre-filtering, silent-feature extraction, endpoint decision, and post-processing. A short steady-state speech signal was divided into multiple frames, each about 20-30 ms long. Depending on the sampling rate

of the audio signal, 256, 512, or 1024 points were taken as a single frame. There were 25%, 50%, and 75% overlaps between adjacent frames. A high-pass filter was used to eliminate low-frequency noise.

### **2.2 Features of recording devices**

Features of recording devices are extracted from audio signals. According to numerous studies [Kraetzer, Oermann, Dittmann et al. (2007b); Aggarwal, Singh, Roul et al. (2014c); Kuresan, Samiappan and Masunda (2019a)], the features of recording devices are generally extracted from non-speech audio segments. Such features include the Fourier coefficient histogram, power spectrum, MFCC, perception linear prediction (PLP), random spectral features (RSFs), bark frequency cepstrum coefficient (BFCC), and linear predictive coding (LPC).

MFCC is the main feature used in speech signal processing. MFCC is resistant to noise, and has high accuracy in speech recognition [Yavuz and Topuz (2018)], speaker recognition, emotion recognition, endpoint detection, and other applications [Aggarwal, Singh, Roul et al. (2014d); Kuresan, Samiappan and Masunda (2019b); Algabri, Bencherif, Alsulaiman et al. (2018); Alali, Dean, Senadji et al. (2017a)]. There is a nonlinear relationship between the frequency of a sound perceived by the human ear and its measured frequency, as defined by Mel scale [Alali, Dean, Senadji et al. (2017b)]. The Mel scale showed the relationship between perceived frequency and measured frequency is linear below 1000 Hz and logarithmic above 1000 Hz. In general, Mel filter banks use 12 or 24 triangular filter banks, and their spectra overlap by 50%. MFCC contains information that is useful for recognizing recording devices and recording environments.

In 2012, Panagakis et al. [Panagakis and Kotropoulos (2012a)] used random spectrum features (RSFs) to extract 325 features from a fixed telephone line and used them as inputs to a sparse representation classifier. The recognition rate of this method reached 95.55% when tested on the Lincoln Labs handset database (LLHDB). This result was better than that of 23 MFCC features. RSF was derived by obtaining the power spectrum of the audio signal and calculating the average power spectrum in the time domain. The RSF of the audio signal was then obtained using the random projection operator.

### **2.3 Algorithm for classifying recording devices**

In general, classification algorithms include Bayes classification algorithm, decision trees, k-nearest neighbor, SRC, logistic regression, SVM [Chen, Xiong, Xu et al. (2019)], Gaussian mixture model (GMM), and neural networks. SVM is most commonly used in recognizing recording devices. SVM uses the classification information of boundary samples and adjusts the discriminant function to conduct pattern recognition on small samples, which are non-linear and have high dimensions. SVM is widely used in endpoint detection [Kumar (2019)], speech recognition [Rajasekhar and Hota (2018)], speaker recognition [Medikonda and Madasu (2018)], recording device recognition [Pandey, Verma and Khanna (2014)], and other applications. The performance of SVM is affected by the penalty factor and kernel function, which can be optimized using genetic algorithm (GA) [Alhroob, Alzyadat, Almukahel et al. (2020)], simulated annealing (SA) [Yeh and Chiang (2017)], or particle swarm optimization (PSO) [Demidova, Nikulchev

and Sokolova (2016)].

In 2014, Kotropoulos et al. [Kotropoulos and Samaras (2014)] applied a radial basis function (RBF) neural network to the recognition of mobile phone from audio signals. The recognition rate reached 97.6% on a mobile phone database, which surpassed SVM and multi-layer perceptron. However, the RBF is limited by the selection of the center for radial basis function, and poor performance with small samples. In recent years, deep learning has been widely used in speech signal processing. In 2012, Abdelhamid et al. [Abdelhamid, Mohamed, Jiang et al. (2012a)] introduced a convolutional neural network (CNN) into the neural network hidden Markov model (NN-HMM) hybrid speech recognition model. Abdelhamid et al. [Abdelhamid, Mohamed, Jiang et al. (2012b)] then used the CNN criterion in the frequency domain to normalize acoustic features, achieving good performance. Mitra et al. [Mitra and Franco (2015)] proposed a time-frequency double CNN for speech recognition. This method outperforms traditional deep neural networks on the Fisher dataset, in terms of noise and background interference and requiring far fewer parameters. Hruz et al. [Hruz and Zajic (2017)] used CNN to detect speaker changes.

#### **2.4 Database of recording devices**

The database of recording devices is divided into fixed telephone audio and mobile phone audio. This database is based on TIMIT [Reynolds (1997a)], HTIMIT, and LLHDB [Reynolds (1997b)], which are accepted by most scholars. The mobile phone audio database has to be constantly updated with the rapid development of mobile phones, telephones, and other electronic products.

**HTIMIT database.** In 1997, Reynolds [Reynolds (1997c)] recorded the audio of 384 speakers (192 men and 192 women) and test signals (such as a 1 Hz scanning signal and Gaussian white noise) from the TIMIT database. Reynolds recorded the speakers and test signals using nine fixed telephones and a microphone, respectively. These recordings were then used to form the HTIMIT database with an 8 KHz sampling rate.

**LLHDB database.** Reynolds [Reynolds (1997d)] recorded the voice of 53 speakers (24 men and 29 women) on nine fixed telephones and a high-quality microphone. Reynolds then used these recordings to construct LLHDB with a sampling rate of 8 KHz.

**Aggarwal audio database.** In 2014, Aggarwal et al. [Aggarwal, Singh, Roul et al. (2014e)] used 26 types of mobile phones from five brands (including Samsung, Sony, and Nokia) to record audio. The audio files were then divided into WAV and AMR formats, and the AMR format was converted into the WAV using FFMPEG software.

According to existing literatures, there are many differences among the components of various recording devices; such components include microphone heads, conditioning circuits, compression algorithms, and speech enhancement algorithms.

### 3 The database

#### 3.1 The recording devices

**Table 1:** Mobile phone models in the database

Name	Model	Name	Model
CC1	Coolpad 5010	CL2	Lenovo A380T
CC2	Coolpad S116	CME1	MEIZU M8
CHA1	Haier N6E	CME2	MEIZU M9
CHU1	Huawei C8500	CMo1	Moto XT502
CHU2	Huawei C5070	CMo2	Moto XT301
CHI1	Hisense E350	CS1	SAMSUNG S579
CHI2	Hisense E89	CS2	SAMSUNG I5700
CK1	K-Touch D8800	CV1	Vivo E1
CK2	K-Touch T619	CZ1	ZTE N600
CL1	Lenovo A66t	CZ2	ZTE N700

The performance of the proposed recognition system was tested on 32 recording devices, spanning 17 brands. The recording devices included 20 mobile phones (shown in Tab. 1.) and 12 recording pens (shown in Tab. 2.). These devices were used to record speech in various environments, such as subway, bus station, library, dormitory, and shopping mall.

**Table 2:** Models of recorders in the database

Name	Model	Name	Model
RJ1	JWD DVR818	RSA1	SAST AY-G30
RJ2	JWD DVR805	RSA2	SAST FX937
RH1	HYM-3698	RT1	TF-A20
RH2	HYM-F97	RT2	TF-a50
RSO1	SONY ICD-fx8	RU1	UNIS V901
RSO2	SONY PCM-M10	RU2	UNIS ZD809

#### 3.2 Conversion of audio file format

Converting compressed audio files into WAV format is the first step in recording device recognition, and it is one of the main challenges in developing recognition systems for recording devices. There are two possible methods for converting audio files into WAV format. The first method involves using existing conversion tools. A variety of conversion tools were found to produce poor results in the form of solidified baud rate and sampling frequency, or relatively complex operation. The decoding algorithms for the various formats of compressed audio files were studied. MP3 and AMR decompression algorithms were considered for converting compressed audio files into WAV format, but the resulting audio was poor. FFMPEG software was finally used for converting the compressed audio files into WAV files. The software had to be recompiled

because its core did not support some compressed audio formats.

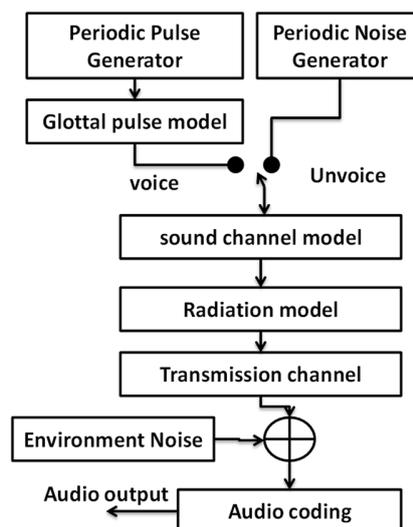


**Figure 1:** Recording devices used for testing the proposed recognition system

#### 4 Feature extraction and the proposed recognition system

Systems for recognizing recording devices from recorded audio vary in terms of their noise spectrum estimation, feature extraction, and pattern recognition schemes. Most audio files come from recording devices in the form of compressed audio files. MP3, AMR, and other compressed audio formats need to be converted into WAV files. The features of recording devices are extracted from speech and noise; therefore, reduction is undertaken on segments of audio files that contain neither. In some literatures [Kraetzer, Oermann, Dittmann et al. (2007c); Abdelhamid, Mohamed, Jiang et al. (2012c)], special attention was paid to extracting features of recording devices from the noise spectrum.

Fig. 3 shows an overview of the recognition system proposed in this article. It is composed of modules for decompression, preprocessing, feature extraction, pattern recognition, amongst others.



**Figure 2:** Model for recording audio signals

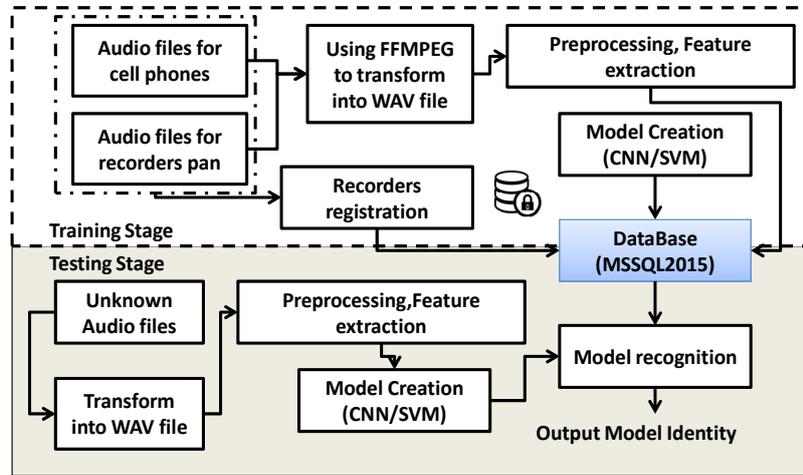


Figure 3: Proposed recognition system

The recording signal  $r(n)$  is given by Eq. (1),

$$r(n) = [s(n) + n(n)] * h(n), \tag{1}$$

where  $s(n)$  is speech signal,  $n(n)$  is environmental signal, and  $h(n)$  is the impulse response of the recording device. The environment noise  $n(n)$  is generally assumed to be Gaussian white noise. Features vary among different recording devices due to differences in the signal acquisition circuits, amplification circuits, noise reduction algorithms, and coding algorithms. These differences are contained in  $h(n)$ . Eq. (2) is given by

$$R(\omega) = (S(\omega) + N(\omega)) \cdot H(\omega). \tag{2}$$

#### 4.1 Preprocessing

The preprocessing involved in recognizing recording devices from recorded audio includes framing, windowing, dynamic speech detection, and noise spectrum estimation. The speech signal was assumed to be stationary for short time-periods, and it was divided into short frames of 20-30 ms. Eq. (3) was derived according to Eq. (2),

$$|H(\omega)|^2 = \frac{P_r(\omega)}{P_s(\omega) + P_n(\omega)}. \tag{3}$$

The single-sided power spectral density function of noise was assumed to be  $n_0$  for noisy frames without speech. Eq. (4) is given by

$$H(\omega) = \sqrt{\frac{P_r(\omega)}{n_0}}. \tag{4}$$

The transfer function of the recording device can be obtained through the power spectrum of a noisy frame.

#### 4.2 Feature extraction

MFCC and RSF are the most effective features for recognizing recording devices from recorded audio. MFCC can sufficiently reduce interference; thus, it is the main feature used in speech and speaker recognition. Studies have not been conducted on whether MFCC can be modified to enhance the recognition of recording devices from recorded audio. Panagakis et al. [Panagakis and Kotropoulos (2012b)] proposed that RSF can obtain higher recording device recognition rates than MFCC. However, RSF requires greater computational complexity due to their higher dimensions.

The research, comparison, and testing of audio files revealed that, above 3400 Hz, the perceived difference between 300 Hz and 3400 Hz of audio was smaller than that between 0 and 300 Hz. RSF in various frequency bands contributes equally to the recognition of recording devices due to their random projection. MFCC in high-frequency bands makes a smaller contribution to the recognition of recording devices, compared to those in low and intermediate frequency bands. Twenty-three MFCCs are not enough to recognize a recording device. Using the following steps, MFCC was improved using the frequency response of the recording device.

##### Framing, windowing, and combination

In each frame of audio signal, the first step was to distinguish noise  $f_{noise}$ , voice  $f_{voice}$ , and unvoice  $f_{unvoice}$ . Threshold was obtained by spectral entropy, as seen in Eq. (5);

$$p_i = Y(\omega_i) / \sum_{k=1}^N Y(\omega_k) \quad (5)$$

$$H = -\sum_{i=1}^N p_i \log p_i, \quad (6)$$

$$sumf_{noise} = \frac{1}{K_1} \sum_{i=0}^{K_1} f_{noise}(i), \quad (7)$$

$$sumf_{voice} = \frac{1}{K_2} \sum_{i=0}^{K_2} f_{voice}(i), \quad (8)$$

$$sumf_{unvoice} = \frac{1}{K_3} \sum_{i=0}^{K_3} f_{unvoice}(i), \quad (9)$$

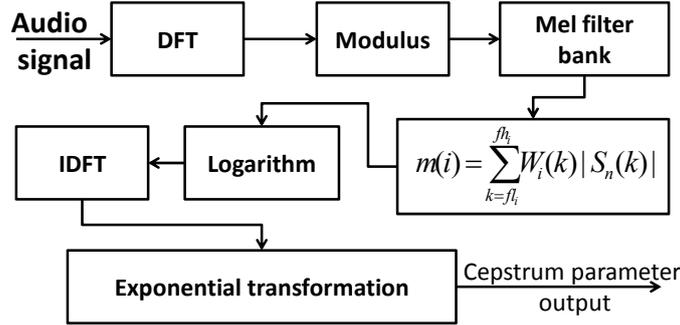
where  $f_{noise}(i)$ ,  $f_{voice}(i)$  and  $f_{unvoice}(i)$  are the noise frames, voice frames, and voiceless frames, respectively.  $sumf_{noise}$ ,  $sumf_{voice}$ , and  $sumf_{unvoice}$  are the mean values of the noise frames, voice frames, and voiceless frames, respectively. Feature extraction is generally done from noise frames, as shown in Fig. 5. The number of noise frames in many audio files is relatively small, which feature extraction difficult. Feature extraction in this paper is done from the noise frames, voice frames, and voiceless frames in order to make the process easier. The sum of audio frames  $S(\omega)$  is expressed as Eq. (10),

$$S(\omega) = \alpha_1 \text{sum}f_{\text{noise}} + \alpha_2 \text{sum}f_{\text{unvoice}} + \alpha_3 \text{sum}f_{\text{voice}}, \quad (10)$$

where  $\alpha_1 + \alpha_2 + \alpha_3 = 1$ , and  $\alpha_1 > \alpha_2 \geq \alpha_3$ .

**Wavelet decomposition**

Speech frames were decomposed with Daubechies 4 (DB4) wavelet to obtain the estimated coefficients in three layers.



**Figure 4: MFCC parameter calculation**

**Calculating Improved MFCC (IMFCC)**

To improve the calculation of high-frequency MFCC, an exponential transform module was added to the process, as seen in Fig. 4.

**4.3 SVM and generalized fractional mixed-Gaussian function**

The SVM used a kernel function to transform low-dimensional space into high-dimensional space to solve the linearly inseparable problems. The SVM then established the optimal classification hyperplane to maximize the separation edge of groups. Eq. (11) is expressed as

$$\begin{cases} \min \frac{\|\omega\|^2}{2} + C \sum_{i=1}^N \xi_i \\ \text{constraint condition } y_i [(\omega \cdot x_i) + b] \geq 1 \end{cases}, \quad (11)$$

where  $x$ ,  $\omega$ , and  $b_i$  are samples, weights, and biases, respectively. The relaxation variable  $\xi_i$  was introduced because of the linearly inseparable case. The SVM generally used the radial basis as its kernel function. Eq. (12) is given by

$$K(x_i, x_j) = \exp(-g \|x_i - x_j\|^2). \quad (12)$$

Results showed that the performance and generalization ability of SVM was affected by the penalty coefficient  $C$  and kernel function coefficient  $g$ . Both of the aforementioned coefficients can be optimized using GA.

The kernel function affects the recognition accuracy of SVM. Gaussian kernel functions are generally used in SVM, but no studies have been conducted on the distribution model

of features of recording devices. Therefore, Gaussian kernel functions do not make for the best models. However, the Gaussian kernel function was extended to the fractional domain to find better options.

An SVM kernel function must satisfy Mercer's theory [Mercer (1909)], put forward by Mercer in 1909. Mercer's theory is an important conclusion of the integral equation theory and a necessary condition for evaluating kernel functions. The commonly used kernel functions are the linear kernel functions, polynomial kernel functions, Gaussian kernel functions, and sigmoid kernel functions.

Theorem 1: For all square integrable functions  $g(x)$ , there is a real function  $K(x, y)$  which satisfies Eq. (13);

$$\iint g(x)K(x, y)g(y)dxdy \geq 0, \quad (13)$$

where  $K(x, y)$  is the kernel function, which can be expressed as a product of mapping function  $\phi(x)$ , as given by Eq. (14);

$$K(x, y) = \phi(x)\phi(y). \quad (14)$$

Using the kernel function to achieve classification, the SVM maps the low-dimensional linearly inseparable case to the high-dimensional linearly separable case. The choice of kernel function affects SVM performance. Using existing kernel functions may not achieve the best classification due to the unpredictable nature of the data being classified. The Gaussian kernel function is the most widely used kernel function in SVM. In this article, a Gaussian kernel function was constructed for classifying the features of a recording device.

The fractional mixed-Gaussian function (FMG) is defined by Eq. (15),

$$K(x, y) = x^\alpha y^\alpha \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right), \quad (15)$$

where  $a$  is the fractional factor. When  $a = 0$ , the kernel function was Gaussian. Furthermore, the FMG was proven to satisfy Mercer's theory. Eq. (16) was obtained by substituting Eq. (15) into Eq. (14);

$$\iint g(x)x^\alpha y^\alpha \exp\left(-\frac{\|y\|^2}{2\sigma^2}\right)\exp\left(-\frac{x^T y}{2\sigma^2}\right)g(y)dxdy. \quad (16)$$

Eq. (17) was obtained by substituting Eq. (13) into Eq. (16). Eq. (18) was then simplified into Eq. (19);

$$\Rightarrow \iint g(x)K(x, y)g(y)dxdy, \quad (17)$$

$$= \iint \sum_{k=0}^{+\infty} \sqrt{\frac{1}{2\sigma^{2k} k!}} (x^T)^k g(x)x^\alpha \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right) \sqrt{\frac{1}{2\sigma^{2k} k!}} y^k g(y)y^\alpha \exp\left(-\frac{\|y\|^2}{2\sigma^2}\right) dxdy, \quad (18)$$

$$= \sum_{k=0}^{+\infty} \iint r(x)r(y)dxdy = \sum_{k=0}^{+\infty} \left(\int r(x)dx\right)^2 \geq 0. \quad (19)$$

Eqs. (16)-(19) demonstrate that Eq. (15) is a kernel function, which satisfies Mercer’s theory. Eq. (15) was, therefore, used as the kernel function of the SVM in this article.

A generalized fractional mixed-Gaussian (GFMG) function was obtained by further extending Eq. (15), as given by Eq. (20);

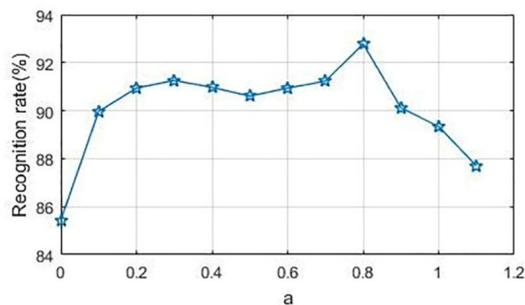
$$K(x, y) = x^\alpha y^\alpha \exp\left(-\frac{\|x - y\|^b}{2\sigma^2}\right). \tag{20}$$

Eq. (20) was a general Gaussian kernel function when  $\alpha=0$  and  $b=2$ .

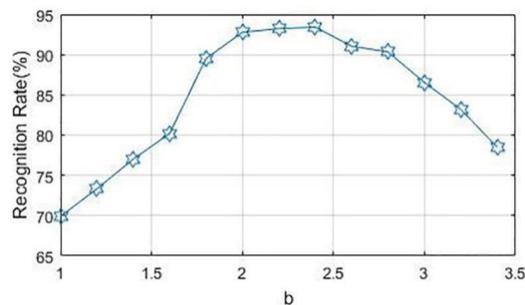
**5 Experiments and discussions**

The audio files used in this experiment were recorded in various environments using the recording devices listed in Tab. 1 and Tab. 2. The sampling frequency was 44.1 KHz and quantization was 8 bits. Each frame length had 2048 points and the frame shift was 50%. Sixty percentage of the audio files were used as training samples and the remaining as test samples.

Figs. 5 and 6 show a plot of the recognition rates of the proposed recognition model versus factors  $\alpha$  and  $b$  under optimal penalty coefficient  $C$  and penalty function coefficient  $g$ . The highest recognition rate in Fig. 5 was 92.78% when  $\alpha=0.8$  and the highest recognition rate in Fig. 6 was 93.42% when  $b=2.4$ . These recognition rates were higher than those obtained when  $\alpha=0$  and  $b=2$ .



**Figure 5:** SVM recognition rate versus  $\alpha$ , when  $b=2$



**Figure 6:** SVM recognition rate vs.  $b$ , when  $\alpha=0.8$

**Table 3:** Performance of various recognition systems

System	RSO2	RSA1	CHU1	CS1	Total
Baseline system (MFCC+SVM)	96.3%	87.9%	78.6%	81.7%	75.8%
IMFCC+SVM	95.9%	91.3%	82.8%	84.1%	87.3%
IMFCC+GFMG+SVM	97.1%	93.2%	84.0%	88.4%	93.4%

Tab. 3 shows the performance of various recognition systems. The system that used IMFCC had higher recognition rates than the baseline system. The proposed system that used IMFCC and GFMG function had the highest recognition rate. Referring to Tab. 3, RSO2 and RSA1 represent recorders from Tab. 2, and CHU1 and CS1 represent mobile phones from Tab. 1. Tab. 3 shows that the recognition rates of recorders were significantly higher than those of mobile phones due to the higher performance of the algorithms and circuits in recorders.

## 6 Conclusions

This article proposed a system for recognizing recording devices from recorded audio. A database of the features of various recording devices was created by using 32 different recording devices to record audio files at different locations. A novel recognition system was created by making improvements to MFCC and using a GFMG function with the SVM model. Experiments revealed that the recognition rates with the GFMG function was higher than those with traditional Gaussian functions.

**Funding Statement:** This work was supported by the Jiangsu University Student Training Program [SJCX19\_0529], the research fund of Nanjing Institute of Engineering [CXY201931], and the National Natural Science Foundation of China (61871213).

**Conflicts of Interest:** The authors declare that they are no conflicts of interest regarding this study.

## References

- Abdelhamid, O.; Mohamed, A.; Jiang, H.; Penn, G.** (2012): Applying convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4277-4280, Kyoto, Japan.
- Aggarwal, R.; Singh, S.; Roul, A. K.; Khanna, N.** (2014): Cellphone identification using noise estimates from recorded audio. *International Conference on Communications and Signal Processing*, pp. 1218-1222, Melmaruvathur, India.
- Alali, A. K. H.; Dean, D.; Senadji, B.; Chandran, V.; Naik, G. R.** (2017): Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions. *IEEE Access*, vol. 5, pp. 15400-15413.

**Algabri, M.; Bencherif, M. A.; Alsulaiman, M.; Muhammad, G.; Mekhtiche, M. A.** (2018): Soft computing techniques for classification of voiced/unvoiced phonemes. *Intelligent Automation and Soft Computing*, vol. 24, no. 2, pp. 267-274.

**Alhroob, A.; Alzyadat, W.; Almukahel, I. H.; Altarawneh, H.** (2020): Missing data prediction using correlation genetic algorithm and SVM approach. *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, pp. 703-709.

**Chen, Y.; Xiong, J.; Xu, W.; Zuo, J.** (2019): A novel online incremental and decremental learning algorithm based on variable support vector machine. *Cluster Computing*, vol. 22, no. 3, pp. 7435-7445.

**Demidova, L.; Nikulchev, E.; Sokolova, Y.** (2016): Big data classification using the SVM classifiers with the modified Particle swarm optimization and the SVM ensembles. *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 5, pp. 294-312.

**Hruz, M.; Zajic, Z.** (2017): Convolutional neural network for speaker change detection in telephone speaker diarization system. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4945-4949, New Orleans, USA.

**Jiang, Y.; Leung, F. H. F.** (2019): Source microphone recognition aided by a kernel-based projection method. *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 11, pp. 2875-2886.

**Kotropoulos, C.; Samaras, S.** (2014): Mobile phone identification using recorded speech signals. *International Conference on Digital Signal Processing*, pp. 586-591, Hong Kong, China.

**Kraetzer, C.; Oermann, A.; Dittmann, J.; Lang, A.** (2007): Digital audio forensics: a first practical evaluation on microphone and environment classification. *Proceedings of the 9th Workshop on Multimedia & Security*, pp. 63-74.

**Kumar, R. K. D. A.** (2019): Single-ended speech quality evaluation using linear combination of the quality score estimates of multi-instances features. *Advances in Electrical and Electronic Engineering*, vol. 12, no. 5, pp. 464-474.

**Kuresan, H.; Samiappan, D.; Masunda, S.** (2019): Fusion of WPT and MFCC feature extraction in Parkinson's disease diagnosis. *Technology and Health Care*, vol. 27, no. 4, pp. 363-372.

**Medikonda, J.; Madasu, H.** (2018): Higher order information set based features for text-independent speaker identification. *International Journal of Speech Technology*, vol. 21, no. 3, pp. 451-461.

**Mercer, J.** (1909): Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, vol. 209, pp. 415-446.

**Mitra, V.; Franco, H.** (2015): Time-frequency convolution networks for robust speech recognition. *Automatic Speech Recognition and Understanding*, pp. 317-323.

**Panagakis, Y.; Kotropoulos, C.** (2012): Automatic telephone handset identification by sparse representation of random spectral features. *Proceedings of the on Multimedia and Security*, pp.91-96, New York, USA.

**Pandey, V.; Verma, V. K.; Khanna, N.** (2014): Cellphone identification from audio recordings using PSD of speech-free regions. *IEEE Students' Conference on Electrical, Electronics and Computer Science*, pp. 1-6, Bhopal, India.

**Rajasekhar, A.; Hota, M. K.** (2018): A study of speech, speaker and emotion recognition using mel-frequency cepstrum coefficients and support vector machines. *International Conference on Communication and Signal Processing*, pp. 0114-0118, Chennai, India.

**Reynolds, D. A.** (1997): HTIMIT and LLHDB: speech corpora for the study of handset transducer effects. *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1535-1538, Munich, Germany.

**Yavuz, E.; Topuz, V.** (2018): A phoneme-based approach for eliminating out-of-vocabulary problem of Turkish speech recognition using hidden Markov model. *International Journal of Computer Systems Science and Engineering*, vol. 33, no. 6, pp. 429-445.

**Yeh, J. P.; Chiang, C. M.** (2017): Reducing the solution of support vector machines using simulated annealing algorithm. *International Conference on Control, Artificial Intelligence, Robotics and Optimization*, pp. 105-108, Prague, Czech Republic.