Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2016) 24: 1942 – 1954

doi:10.3906/elk-1312-193

Identifying acquisition devices from recorded speech signals using wavelet-based features

Ömer ESKİDERE*
Department of Electrical-Electronics Engineering, Faculty of Engineering, Bursa Orhangazi University, Yıldırım, Bursa, Turkey

Received: 24.12.2013 • **Accepted/Published Online:** 01.07.2014 • **Final Version:** 23.03.2016

Abstract: Speech characteristics have played a critical role in media forensics, particularly in the investigation of evidence. This study proposes two wavelet-based feature extraction methods for the identification of acquisition devices from recorded speech. These methods are discrete wavelet-based coefficients (DWBCs) and wavelet packet-based coefficients, which are mainly based on a multiresolution analysis. These features' ability to capture characteristics of acquisition devices is compared to conventional mel frequency cepstral coefficients and subband-based coefficients. In the experiments, 14 different audio acquisition devices were trained and tested using support vector machines. Experimental results showed that DWBCs can effectively be used in source audio acquisition device identification problems.

Key words: Media forensics, acquisition device identification, wavelet packet transform, discrete wavelet transform

1. Introduction

Speech signal processing, which involves information contained in the speech signal, has received attention in recent years. A speech signal comprises different forms of information such as the conveyed message; the identity, emotion, age, and gender of the speaker; and information about the recording device [1–5].

Today, a speech signal can be doctored easily by using audio editing software. For media forensics, it is an important problem to verify the integrity of speech signals [6–8]. To tackle this problem, this study addresses the task of automatic acquisition device detection. In particular, information about an acquisition device from speech records may be of great importance in criminal investigations of evidence. Furthermore, a source acquisition device has unique traces related to the speech record pipeline of the device, which can help forensic examiners in trying to verify the integrity of the content. For the evaluation of evidence, the acquisition device can be used to determine whether the speech signal has been manipulated or tampered with. Therefore, the automatic detection of a source acquisition device has become essential, especially for law enforcement.

In the literature, a small amount of research has been done in determining acquisition devices based on speech recordings. Romero and Wilson [9] investigated the performance of mel frequency cepstral coefficients (MFCCs) by using a support vector machine (SVM) classifier for source acquisition device identification. In their study, classification accuracies of higher than 90% for landline telephone handsets and microphones were achieved. In our previous work [5], we proposed a novel idea based on a MFCC feature extraction technique for source cell phone identification. The vector quantization and a SVM framework based on a generalized linear discriminant sequence kernel were employed. The best correct cell phone identification results were achieved as

*Correspondence: omer.eskidere@bou.edu.tr

96.42% by the SVM classifier on a set of 14 models of cell phones. In addition, several other similar works [10–13] also exist in the area of source microphone classification, which could be considered similar to the identification of acquisition devices.

All of these studies used standard feature extraction and pattern recognition techniques, which are taken from work previously done in the fields of speech recognition and speaker recognition. Usually, MFCCs have been used to date as fundamental speech features in acquisition device recognition systems. MFCCs are based on the model of humans' auditory perception and have been proven to be very effective as a speech feature extraction technique in a variety of problems [14–16]. Widespread uses of MFCCs for recognition tasks arise from their ability to represent the speech spectrum in a compact form [17]. However, MFCCs are not immune to environmental noise and degradation in recognition rates [18].

A popular approach in signal processing is the wavelet transform, which has the ability to allow simultaneous time and frequency analysis and has replaced Fourier analysis in many applications [19–22]. The wavelet transform is a suitable technique since it can obtain important speech features that are robust against noise as well as minor variations [23–25]. For speech signal analysis, this method uses a variable window to scan the frequency spectrum, increasing the temporal resolution of the analysis. There are different wavelet-based features that were reported to outperform MFCCs, such as wavelet-based MFCCs [24], multiband linear predictive cepstral coefficients [25], and subband cepstral coefficients and wavelet packet parameters [26]. These works indicate that wavelet-based features can successfully be employed for speaker identification problems.

The aim of this study is to present a novel automatic acquisition device detection method based on discrete wavelet and wavelet packet decomposition, and an SVM classifier. Usually, it is difficult to understand the acquisition device from recorded speech signals by listening tests for an untrained general user or inexperienced forensic examiner. Because the characteristics of acquisition devices are different from each other, the proposed decomposition methods are used for decomposing more detailed information of speech signals. In the decision-making phase from the obtained features, the SVM classifier was used and tested to validate the applicability and efficiency of the acquisition device detection method with discrete wavelet-based coefficients (DWBCs)/wavelet packet-based coefficients (WPBCs). The proposed wavelet-based features with MFCCs and subband-based coefficients (SBCs) were compared to determine the appropriate features. The results showed that the proposed methods provided a high detection rate compared to the other features.

2. The acquisition device identification system

The acquisition device identification system consists of three stages: source recording, feature extraction, and classification. Source recording is the first step in the method. This step involves taking speech samples from different acquisition devices and loading them onto a computer. This huge data set is then reduced by using feature extraction techniques. Using these techniques, feature vectors are derived, which uniquely characterize the information. In the third step, a classifier is used for classification of the extracted features from the previous step. For this, a SVM classifier is used.

2.1. Source recording

In the source recording, 14 different models of portable acquisition devices for speech recordings were collected. The brands and models of these acquisition devices are shown in Table 1. These devices, which incorporate a built-in microphone, were chosen since they allow using the same recording format (wav) and are widely used. The software, called Audacity, is used for recordings on portable computers (notebook computer, netbook

computer, ultrabook computer, and tablet PC). For smart phones, different software, Hertz, is used to obtain wav recordings. These recordings were then transmitted to a PC and were fixed at a sampling rate of 16 kHz by Audacity with a mono channel.

Table 1. The brands and models of portable acquisition devices used in the experiments.

Id	Acquisition device	Brand-model
D1	Mp3 player	Creative-Zen Mozaic
D2	Ultrabook computer	Dell-Xps13
D3	Netbook computer	HP-Compaq mini
D4	Smart phone	HTC-Wildfire S
D5	Smart phone	LG-Optimus
D6	Tablet PC	Onyo-Powerpad
D7	Mp3 player	Samsung-YPU2
D8	Smart phone	Samsung-Galaxy S2
D9	Smart phone	Samsung-Galaxy Wonder
D10	Mp3 player	Sony-NWZ B152
D11	Digital voice recorder	Sony-ICD Ux512
D12	Smart phone	Sony-Xperia
D13	Notebook computer	Toshiba-Satellite
D14	Digital voice recorder	Zoom-H4N

In the identification of acquisition devices, speech files were created by using the TIMIT database. For each acquisition device, speech utterances of 12 min in length were recorded in a silent room from 24 speakers (12 of each gender). Thus, the experiment involved 336 speakers in total.

2.2. Wavelet analysis

For nonstationary signals such as human speech, signal frequencies' components vary with time. In order to analyze these signals, the short-time Fourier transform (STFT) is used. The STFT can be considered as the filter bank and consists of finite impulse response filters with equal frequency resolution. In this method, it is difficult to meet sharp localization in time and frequency simultaneously. Therefore, the STFT is not always an optimal method for analyzing speech signals. This difficulty can be handled using multiresolution time-frequency analysis approaches. A multiresolution representation of a signal can be obtained by using the discrete wavelet analysis. In that case, the discrete wavelet transform (DWT) coefficients of a continuous time signal, $s(t)$ are represented by

$$W_{a,b} = \int_{-\infty}^{+\infty} s(t)\ \psi_{a,b}(t)\, dt \tag{1}$$

where a ($a \neq 0$) and b are dilation (scaling) and translation (shifting) parameters, respectively. The DWT basis $\psi_{m,n}(t)$ can be given by

$$\psi_{m,n}(t) = a_0^{-m/2}\, \psi\left(a_0^{-m} t - n b_0\right), \tag{2}$$

where discretizations of parameters are $a = a_0^m$ and $b = nb_0a_0^m$, respectively. Discrete wavelet coefficients $(W_{m,n})$ are expressed as

$$W_{m,n} = \int_{-\infty}^{+\infty} s(t)\ \psi_{m,n}(t)\, dt \tag{3}$$

In discrete wavelet analysis, a signal is decomposed into one containing low-frequency information and another containing high-frequency information [22]. In other words, a signal is split into an approximation component and a detail component. Next, the approximation component is split into a second-level approximation component and detail component and this process can be repeated iteratively. In this way, the speech signal is decomposed into several frequency subbands. The number of subbands depends on the decomposition level. For a signal of $j-$level decomposition, $j+1$ possible ways of decomposition can be obtained.

Wavelet packet analysis is an extension of the DWT, which provides a richer range of possibilities for signal analysis. In wavelet packet analysis, the signal is decomposed into both approximations and details result in a balanced tree structure. For a given level j, the wavelet packet transform (WPT) decomposes the signal into 2^j subbands. In this way, the optimum time-frequency representation of the speech signal can be obtained [27]. Figure 1 shows the WPT decomposition and DWT decomposition tree structures for three levels.

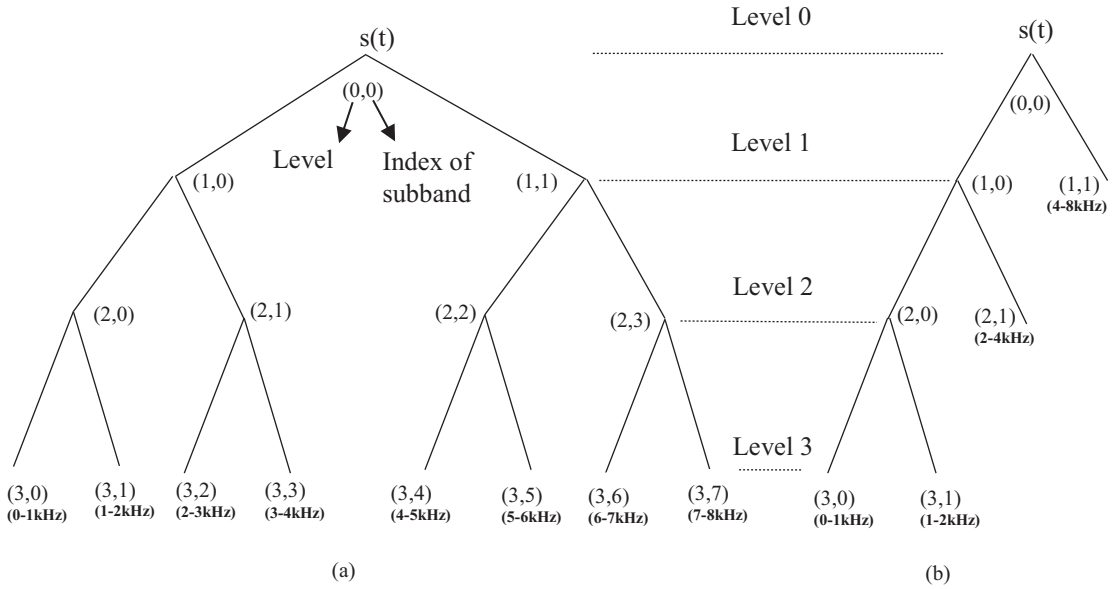

Figure 1. The tree structure of the three-level (a) WPT decomposition and (b) DWT decomposition.

2.3. DWBC and WPBC features

In this subsection, two new sets of features based on DWT and WPT for acquisition device detection are presented. The block diagram of the proposed speech features, DWBCs and WPBCs, is illustrated in Figure 2.

First, a speech signal $s(n)$ is divided into overlapping frames, each of length 32 ms with a skip rate of 16 ms. For WPBCs, wavelet packet decomposition is applied for each speech frame. In the second method, for DWBCs, DWT decomposition is applied and subband signals are obtained. The optimal number of decomposition levels can be determined experimentally. Both WPBCs and DWBCs are implemented by using the Daubechies wavelet filter. Linear prediction coefficients (LPCs) are then estimated from these subband

signals. For a subband signal having N samples, $\{s_1, s_2, \ldots, s_N\}$, linear prediction analysis is based on the assumption that each speech sample is approximately predicted by a linear combination of α past samples [28]:

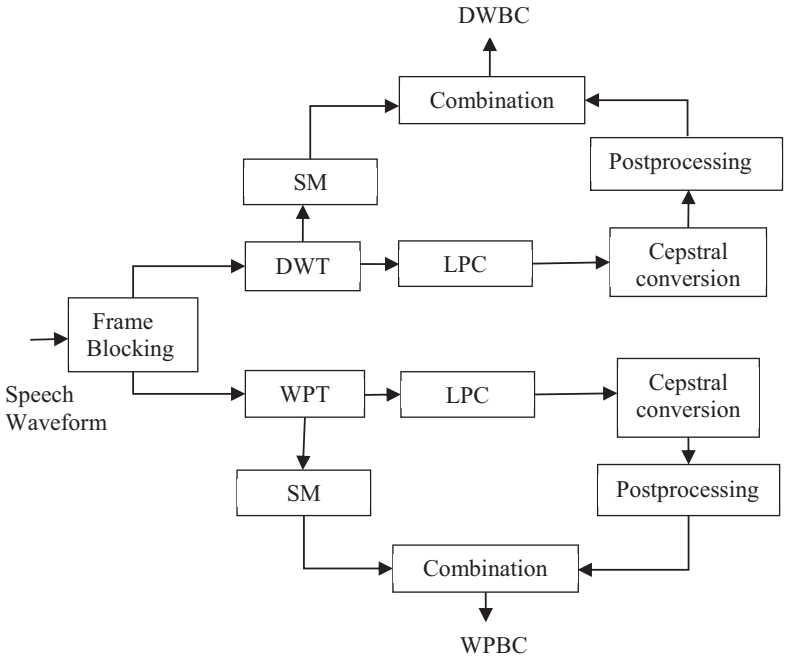

Figure 2. Block diagram of DWBC and WPBC feature extraction methods.

$$\hat{s}(n) = -\sum_{k=1}^{\alpha} a_k s(n-k) \tag{4}$$

where α is the order of linear prediction analysis and $\{a_k\}_{k=1}^{\alpha}$ are the linear prediction coefficients. The suitable value of α can be derived from subbands experimentally. The LPCs can be computed by minimizing the energy of the prediction residual.

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{\alpha} a_k s(n-k) \tag{5}$$

These linear prediction coefficients can be estimated by the autocorrelation method, which uses the Levinson–Durbin algorithm.

The LPCs are then converted to cepstral coefficients from a_k coefficients through the following equations:

$$c_1 = a_1, \tag{6}$$

$$c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}, \quad 1 < m < \alpha \tag{7}$$

$$c_m = \sum_{k=m-\alpha}^{m-1} \frac{k}{m} c_k a_{m-k}, \quad m > \alpha \tag{8}$$

In the postprocessing step, linear prediction cepstral coefficients (LPCCs) are combined from every subsignal and first-order coefficients are discarded. The main reason for using these parameters is their good representation of the envelope of the spectrum of utterances.

For proposed features, statistical measures (SMs) of subband signals and LPCC features can be combined together to produce a system that performs good feature extraction. SMs of each subband signal can provide complementary information to the LPCC of each subband. Different SMs such as various entropies (Shannon entropy, log-energy entropy, and sure entropy), the standard deviation (STD), and the mean can be employed. In particular, entropy can be defined as the statistical measure of information or uncertainty [28]. The log-energy entropy is computed for each subband signal as

$$E_{le}(s) = \sum_{m=1}^{N} log(s_m^2) \tag{9}$$

Shannon entropy:

$$E_{sh}(s) = -\sum_{m=1}^{N} s_m^2 log(s_m^2) \tag{10}$$

Sure entropy:

$$E_{su}(s) = n - \#\{\, i \,\text{such that}\, |s_m| \leq p + \sum_{m=1}^{N} min(s_m^2, p^2) \tag{11}$$

Mean:

$$M(s) = \frac{1}{N} \sum_{m=1}^{N} s_m \tag{12}$$

Standard deviation:

$$STD(s) = \left(\frac{1}{N-1} \sum_{m=1}^{N} (s_m - \bar{s})^2 \right)^{\frac{1}{2}}, \; where \; \bar{s} = \frac{1}{N} \sum_{m=1}^{N} s_m \tag{13}$$

Here, N is the number of coefficients in the subband, s_m are the subband signals, and p is a positive threshold. To form a single feature vector, cepstral coefficients and statistical measures from each subband are combined and finally the DWBC/WPBC is obtained.

3. Parameter settings

In parameter settings, all speech features were extracted from the speech signal every 16 ms using a 32-ms Hamming window. The same speech frame size for comparison of features was kept in all experiments. For comparison, proposed wavelet-based features with MFCCs and SBCs are used. MFCCs are the most widely used speech features currently, while SBCs are also used in some speech recognition and speaker recognition applications [17,26]. In the MFCC extraction, there are various implementations [29]; a method that was proposed by Slaney [30] was selected here due to its good performance. One of the WPT applications is wavelet subband-based analysis [26]. Using this method, SBCs, which use perceptual wavelet packet decompositions, can be obtained. First, for each speech frame, wavelet packet decomposition is applied. The tree structure of

the wavelet packet filter bank is chosen, which especially emphasizes low and middle frequencies. Second, the subband energies are computed at the output of the wavelet packet filter bank. Finally, SBCs are derived by using the discrete cosine transform from the log-subband energies. Detailed information about this technique can be found in [17,26]. For the experiments, a wavelet packet decomposition of the frequency ranges [0–8 kHz] and 24 frequency subbands were used.

For evaluation settings, each data set was divided into two parts as training and testing data. Half of the data were used as training data (6 min for each device) and the rest were kept for testing. In the testing stage, all recorded data were segmented into chunks of 3 s in length [5] and the total number of test data used in the data set was 1680 (120 tests for each acquisition device) for all recording conditions. The Lib-SVM package has been used in the experiments, which uses a one-versus-one multiclass classification approach [31]. Moreover, an RBF kernel width parameter σ and penalty parameter C of the SVM classifier were chosen experimentally as $\sigma = 2$ and $C = 5$, respectively.

4. Experimental results

In this section, the speech signals were tested to evaluate the proposed acquisition device detection method. The optimal parameters of the proposed system were investigated, including the type of SMs of the subband signals, decomposition level of DWBC/WPBC, and order of linear prediction analysis used to calculate LPCs of each subband.

In the first experiment, suitable approaches of SMs of the subband for DWBCs and WPBCs were investigated for acquisition device identification performance. For this purpose, Shannon entropy, sure entropy, log-energy entropy, standard deviation, and mean of subband signals were tested. Table 2 denotes the acquisition device identification performance with various SMs of subband signals. Parameters of the features were set by using fifth-order linear prediction analysis for each subband, and the decomposition level of DWBCs and WPBCs was set as 3 and 2, respectively. We analyzed the performance of the baseline device identification system without any statistical measure and this method is referred to as the baseline in Table 2. In this case, features obtained are only cepstral coefficients.

Table 2. Acquisition device identification accuracy (%) for various statistical measures (SMs) of subband signals using DWBCs and WPBCs.

SM	DWBC	WPBC
Baseline without SM	72.56	70.30
Shannon entropy	62.56	67.38
Sure entropy (p = 3)	79.29	77.50
Log-energy entropy	81.96	77.68
Mean	72.56	77.86
STD	82.44	80.24

As seen in Table 2, SMs including log-energy entropy, standard deviation, and sure entropy improve the acquisition device identification accuracy considerably in comparison to the baseline system (without any SMs) for DWBCs. Comparing two types of speech features, DWBC outperforms WPBC and the acquisition device identification results obtained with the DWBC feature is 82.44% for the STD and 81.96% for log-energy entropy (chance level is 1/14, thus 7.14%). This might be because of the fact that these SMs give better device-specific information from subbands than other statistical measures.

Next, we evaluated the impact of various decomposition levels for the proposed feature extraction methods. In this case, different orders of linear prediction analysis (varying from 5 to 25) of subband signals were employed. Identification rates using DWBCs with log-energy entropy and the STD of each subband signal are given in Figure 3.

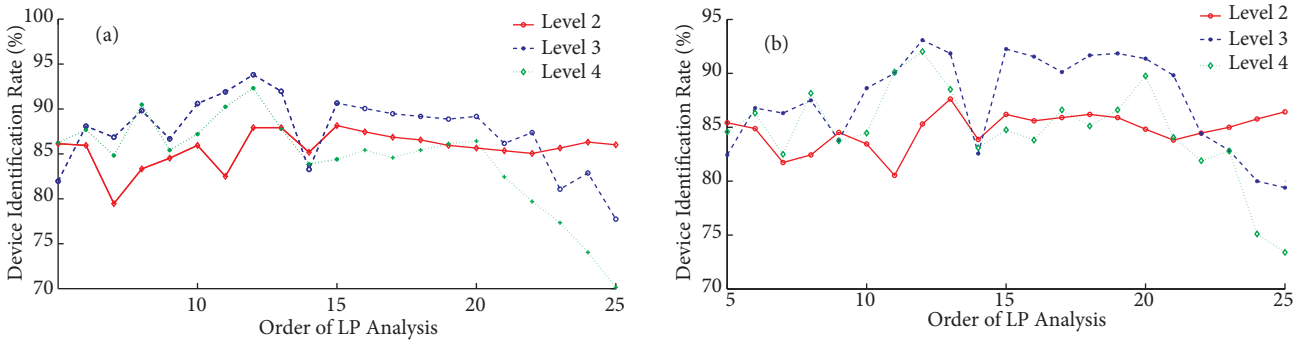

Figure 3. Acquisition device identification rate versus order of linear prediction (LP) analysis for DWBCs using SMs: (a) log-energy entropy and (b) standard deviation.

As shown in Figure 3, the acquisition device accuracy is significantly affected by the selected decomposition levels for both SMs. The best identification rates of 93.81% for log-energy entropy and 93.08% for STD were achieved when the decomposition level was equal to 3. WPBC features were also evaluated for various decomposition levels and order of linear prediction analysis. Identification rates are presented in Figure 4.

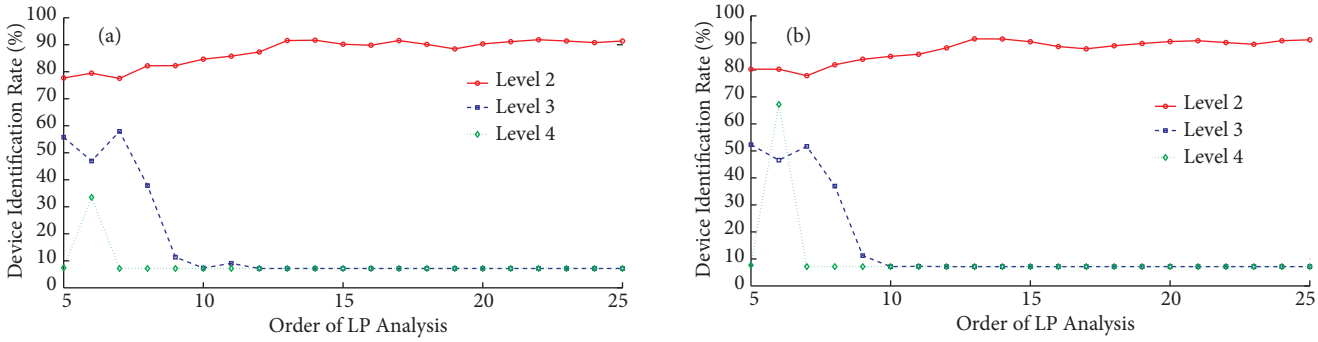

Figure 4. Acquisition device identification rates versus order of linear prediction (LP) analysis for WPBCs using SMs: (a) log-energy entropy and (b) standard deviation.

From Figure 4, it can be seen that two-level decomposition outperformed three- and four-level decomposition in detecting devices. With increases in decomposition levels (more subbands), more information can be obtained from the speech signal. On the other hand, this causes increases in the number of useless features, as well. In the third decomposition level, useless features further increase, leading to multiclass decisions toward a single class for all estimated classes. With decomposition level four, WPBC features failed in identifying an acquisition device as indicated by an identification accuracy of 7.14% because of the excessive useless features. As can be deduced from Figures 3 and 4, DWBCs usually have superior performance compared to WPBCs.

Figures 3 and 4 reveal that the recognition rates of devices were affected by the order of linear prediction. Generally, the optimal number of order of linear prediction analysis was 12, above and below which the

performance slowly degraded for DWBC features. The 14th order linear prediction analysis gives the best recognition rates for WPBCs.

The best result obtained is illustrated with the confusion matrix of DWBC features discussed below. These features are derived from three levels of decomposition of the wavelet transform with the SM, log-energy entropy, and the order of linear prediction analysis is set as 12. Confusion tables for DWBC features are demonstrated in Table 3, and the correct decisions (diagonal elements) are marked in bold. From Table 3, we can see that the largest confusions occur between *D11* and *D9*, and between *D9* and *D6* (they are represented in gray). On the other hand, *D1* and *D13* are the best classified acquisition devices, while *D9* and *D11* are poorly classified by DWBCs. An interesting observation is that the quite similar smart phones from the same manufacturer (Samsung-Galaxy S2 and Samsung-Galaxy Wonder) get mixed up even less often than is the case with other device combinations. The anomaly is the frequent misclassification of the Samsung-Galaxy Wonder (D9) as the Onyo-Powerpad (D6). This asymmetric error may be attributed to these two devices sharing the same transducer technology. Furthermore, the other reason could be feature types. Additional features may be used to increase the discriminatory power of the features.

Table 3. Confusion matrix of acquisition device identification system for DWBCs.

	D1	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*	*D8*	*D9*	*D10*	*D11*	*D12*	*D13*	*D14*
D1	**120**	0	0	0	0	0	0	0	0	0	0	0	0	0
D2	0	**112**	3	0	3	2	0	0	0	0	0	0	0	0
D3	0	0	**119**	0	0	0	0	0	1	0	0	0	0	0
D4	0	0	0	**117**	0	0	0	3	0	0	0	0	0	0
D5	0	0	0	0	**107**	6	0	0	0	0	0	3	0	4
D6	0	0	0	0	1	**115**	0	0	0	3	0	0	0	1
D7	0	0	0	0	0	0	**112**	0	0	0	7	1	0	0
D8	0	0	0	1	0	0	0	**119**	0	0	0	0	0	0
D9	0	0	0	0	0	12	0	1	**104**	0	0	3	0	0
D10	0	0	0	0	0	5	0	0	0	**109**	6	0	0	0
D11	0	0	0	0	0	0	0	0	14	0	**106**	0	0	0
D12	0	1	0	0	0	5	0	0	0	1	0	**109**	0	4
D13	0	0	0	0	0	0	0	0	0	0	0	0	**120**	0
D14	0	0	0	0	0	5	0	0	0	1	0	5	2	**107**
Accuracy = 93.81% (1576/1680)														

Table 4 presents a comparison among the wavelet-based features and MFCCs. For a fair comparison, the same parameters and training and testing data were used for all speech features. MFCCs with 18 and 24 dimensions were computed and delta coefficients Δ were also added. WPBC and DWBC features were derived using the STD measurement of each subband, and order of linear prediction analysis was set as 5. The level of wavelet transform was set as 2 for 18 DWBCs and 24 WPBCs. In addition, three levels of DWT were performed for 24 DWBCs. Acquisition device identification performances for DWBCs, WPBCs, SBCs, and MFCCs are given in Table 4.

From Table 4, DWBC features are clearly seen to outperform the MFCC, MFCC+Δ, SBC, and WPBC features. Due to better representation of the device-specific variations in speech, DWBC features illustrated a superior performance compared to the others. After DWBCs, the best correct identification result was obtained by using 24 MFCCs. On the other hand, the use of delta MFCCs decreases the performance of device detection systems. Hence, the results of MFCCs and DWBCs in Table 4 can be best represented by the confusion matrices in Tables 5 and 6.

Table 4. Comparison between the proposed features and other approaches.

Features	Dimensions	Accuracy (%)
MFCC	18	75.30
	24	80.65
MFCC+Δ	24 (12+12)	68.99
MFCC+Δ	48 (24+24)	78.99
SBC	24	59.58
WPBC	24	80.24
DWBC	18	85.42
	24	82.44

Table 5. Confusion matrix of acquisition device identification system for 18 dimensional MFCCs.

	D1	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*	*D8*	*D9*	*D10*	*D11*	*D12*	*D13*	*D14*
*D*1	**108**	0	0	7	0	0	0	0	5	0	0	0	0	0
*D*2	0	**117**	3	0	0	0	0	0	0	0	0	0	0	0
*D*3	0	0	**120**	0	0	0	0	0	0	0	0	0	0	0
*D*4	25	0	0	**95**	0	0	0	0	0	0	0	0	0	0
*D*5	2	1	0	0	**24**	0	7	6	12	0	19	49	0	0
*D*6	0	39	0	0	0	**35**	6	0	1	38	0	0	0	1
*D*7	0	3	0	0	0	0	**117**	0	0	0	0	0	0	0
*D*8	14	2	0	31	0	0	0	**71**	1	0	0	1	0	0
*D*9	1	6	0	1	0	0	8	0	**89**	0	15	0	0	0
*D*10	0	2	0	0	0	0	0	0	0	**118**	0	0	0	0
*D*11	1	1	0	0	1	0	12	0	47	0	**58**	0	0	0
*D*12	0	0	0	1	18	3	1	2	2	1	1	**87**	0	4
*D*13	0	0	0	0	0	0	0	0	0	0	0	0	**120**	0
*D*14	0	10	0	0	0	0	1	0	0	3	0	0	0	**106**
Accuracy = 75.29% (1265/1680)														

Table 6. Confusion matrix of acquisition device identification system for 18 dimensional DWBCs.

	D1	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*	*D8*	*D9*	*D10*	*D11*	*D12*	*D13*	*D14*
*D*1	**116**	0	4	0	0	0	0	0	0	0	0	0	0	0
*D*2	0	**95**	6	0	11	4	0	0	2	0	0	0	0	2
*D*3	0	0	**115**	0	0	0	0	0	3	1	0	0	0	1
*D*4	0	0	0	102	0	0	0	15	2	0	0	1	0	0
*D*5	0	6	0	0	**98**	4	0	0	0	0	0	10	0	2
*D*6	0	5	1	0	5	**102**	1	0	0	4	0	0	0	2
*D*7	0	0	0	0	0	12	**105**	0	0	0	3	0	0	0
*D*8	0	0	0	0	0	0	0	**120**	0	0	0	0	0	0
*D*9	0	0	0	1	2	22	0	0	**91**	3	0	1	0	0
*D*10	0	1	0	0	0	1	0	0	0	**115**	3	0	0	0
*D*11	0	0	0	0	0	1	2	0	5	19	**93**	0	0	0
*D*12	1	6	0	0	23	29	0	0	2	0	0	**53**	0	6
*D*13	0	0	2	0	0	0	0	0	0	0	0	0	**118**	0
*D*14	0	0	0	0	2	6	0	0	0	0	0	0	0	**112**
Accuracy = 85.42% (1435/1680)														

From Tables 5 and 6, it can be observed that the classification accuracy of 75.29%, being the worst performance, is for *D5* (which was often confused with *D12*). The DWBC-based system presents classification accuracy of 85.42%, with *D12* and *D9* being the acquisition devices posing the biggest challenges.

5. Discussion

In an attempt to find an efficient representation of speech signals for the determination of acquisition devices, we seek alternative ways to depict the acquisition device's uniqueness. In order to achieve that, DWBC and WPBC feature extraction techniques are proposed. The experiments showed that the proposed features were effective in representing the characteristics of acquisition devices. However, the optimal parameters of these features, such as type of SMs, order of linear prediction analysis, and decomposition level, need to be set. They are important factors for the success of acquisition device identification systems. Table 3 shows that there is much distinguishing acquisition device information present in the DWBCs, which by themselves produced an acquisition device identification rate of 93.81%. The results in Table 4 show that the proposed approach (DWBC) could achieve more favorable results compared to MFCCs and SBCs because DWBC features allow capturing additional information regarding the acquisition device patterns.

We also compared results obtained in this study to the existing results reported in the literature. Acquisition device recognition performances by various researchers are depicted in Table 5. In [5], Hanilçi et al. considered 12 dimensions of MFCCs along with their first-order derivatives as the feature set for source cell phone recognition problems. Romero and Wilson [9] used 23 MFCCs for each frame in their experiments. Based on these results, comparison to other systems is not reasonable because of the different types of devices or different methods of performance evaluation retained.

6. Conclusions

In this study, new effective features based on wavelet transform are presented with applications to acquisition device identification. These features, namely DWBCs and WPBCs, were derived. The experimental results showed that the proposed DWBC technique is able to outperform MFCC, SBC, and WPBC features for 14 different acquisition devices. The main advantage of this system is that it can better represent speech record pipelines of devices by using DWBCs. While the proposed system generally leads to good results, it highly depends on the decomposition level of wavelet transform, type of SMs of the subband signals, and order of linear prediction analysis. Future work is recommended for the extension of these methods to manipulation detection problems.

References

[1] Reynolds DA. Speaker identification and verification using Gaussian mixture speaker models. Speech Commun 1995; 17: 91-108.

[2] Rabiner LR, Juang BH. Fundamentals of Speech Recognition. Englewood Cliffs, NJ, USA: Prentice Hall, 1993.

[3] Chul ML, Narayanan S. Toward detecting emotions in spoken dialogs. IEEE T Speech Audi P 2005; 13: 293-303.

[4] Metze F, Ajmera J, Englert R, Bub U, Burkhardt F, Stegmann J, Muller C, Huber R, Andrassy B, Bauer JG et al. Comparison of four approaches to age and gender recognition for telephone applications. In: Proceedings of the ICASSP; 15–20 April 2007; Honolulu, HI, USA. New York, NY, USA: IEEE. pp. 1605-1608.

[5] Hanilçi C, Ertaş F, Ertaş T, Eskidere Ö. Recognition of brand and models of cell-phones from recorded speech signals. IEEE T Inf Foren Sec 2012; 7: 625-634.

[6] Grigoras C. Applications of ENF criterion in forensic audio, video, computer, and telecommunication analysis. Forensic Sci Int 2007; 167: 136-145.

[7] Nicolalde DP, Apolinário JA, Biscainho LWP. Audio authenticity: detecting ENF discontinuity with high precision phase analysis. IEEE T Inf Foren Sec 2010; 5: 534-543.

[8] Yang R, Zhenhua Q, Jiwu H. Detecting digital audio forgeries by checking frame offsets. In: Proceedings of MM&Sec'2008; 22–23 August 2008; Oxford, UK. New York, NY, USA: ACM. pp. 21-26.

[9] Romero DG, Wilson CYE. Automatic acquisition device identification from speech recordings. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing; 14–19 March 2010; Dallas, TX, USA. New York, NY, USA: IEEE. pp. 1806-1809.

[10] Kraetzer C, Oermann A, Dittmann J, Lang A. Digital audio forensics: a first practical evaluation on microphone and environment classification. In: 9th Workshop on Multimedia & Security; 2007. New York, NY, USA: ACM. pp. 63-74.

[11] Buchholz R, Kraetzer C, Dittmann J. Microphone classification using Fourier coefficients. Lect Notes Comp Sci 2009; 5806: 235-246.

[12] Kraetzer C, Schott M, Dittmann J. Unweighted fusion in microphone forensics using a decision tree and linear logistic regression models. In: Proceedings of the 11th Workshop on Multimedia and Security; 2009. Princeton, NJ, USA: ACM Press. pp. 49-56.

[13] Kraetzer C, Qian K, Schott M, Dittmann J. A context model for microphone forensics and its application in evaluations. In: Proceedings of Media Watermarking, Security, and Forensics XIII, Electronic Imaging Conference; 2011. New York, NY, USA: SPIE.

[14] Tzanetakis G, Cook P. Musical genre classification of audio signals. IEEE T Speech Audi P 2002; 10: 293-301.

[15] Cho HY, Oh YH. On the use of channel-attentive MFCC for robust recognition of partially corrupted speech. IEEE Signal Proc Let 2004; 11: 581-584.

[16] Campbell WM, Campbell JP, Reynolds DA, Singer E, Torres-Carrasquillo PA. Support vector machines for speaker and language recognition. Comput Speech Lang 2006; 20: 210-229.

[17] Sarikaya R, Hansen HL. High resolution speech feature parameterization for monophone-based stressed speech recognition. IEEE Signal Proc Let 2000; 7: 182-185.

[18] Erzin E, Cetin AE, Yardımci Y. Subband analysis for robust speech recognition in the presence of car noise. In: Proceedings of ICASSP-95; 9–12 May 1995; Detroit, MI, USA. New York, NY, USA: IEEE. pp. 417-420.

[19] Phadke AG, Thorp JS. Computer Relaying for Power Systems. 2nd ed. Baldock, UK: Research Studies Press Ltd., 2009.

[20] Garcia C, Zikos G, Tziritas G. A wavelet-based framework for face recognition. In: International Workshop on Advances in Facial Image Analysis Recognition Technology; 1998.

[21] Mallat S. A theory for multiresolution signal decomposition: the wavelet representation. IEEE T Pattern Anal 1989; 11: 674-693.

[22] Mallat S. A Wavelet Tour of Signal Processing. San Diego, CA, USA: Academic Press, 1998.

[23] Tufekci Z, Gurbuz S. Noise robust speaker verification using mel-frequency discrete wavelet coefficients and parallel model compensation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing; 18–23 March 2005. New York, NY, USA: IEEE. pp. 657-660.

[24] Mahmoud IA, Hanaa SA. Wavelet-based mel-frequency cepstral coefficients for speaker identification using hidden Markov models. J Telecommun 2010; 1: 16-21.

[25] Chen WC, Hsieh CT, Lai E. Multiband approach to robust text-independent speaker identification. Computational Linguistics and Chinese Language Processing 2004; 9: 63-76.

[26] Sarikaya R, Pellom BL, Hansen HL. Wavelet packet transform features with application to speaker identification. In: Proceedings of the IEEE Nordic Signal Processing Symposium; 1998; Visgo, Denmark. New York, NY, USA: IEEE. pp. 81-84.

[27] Keeton PIJ, Schlindwein FS. Application of wavelets in Doppler ultrasound. Sensor Rev 1997; 17: 38-45.

[28] Campbell JP. Speaker recognition: a tutorial. P IEEE 1997; 85: 1437-1462.

[29] Ganchev T, Fakotakis N, Kokkinakis G. Comparative evaluation of various MFCC implementations on the speaker verification task. In: Proceedings of the SPECOM; 2005. pp. 191-194.

[30] Slaney M. Auditory Toolbox: A MATLAB Toolbox for Auditory Modeling, Work Technical Report. Palo Alto, CA, USA: Interval Research Corporation, 1998.

[31] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM T Intel Syst Tec 2001; 2: 1-27.