ISSN 0146-4116, Automatic Control and Computer Sciences, 2019, Vol. 53, No. 1, pp. 72–79. © Allerton Press, Inc., 2019.

Detection of HMM Synthesized Speech by Wavelet Logarithmic Spectrum[1]

Diqun Yan[a, b, *], Li Xiang[a], Zhifeng Wang[a], and Rangding Wang[a]

[a] *College of Information Science and Engineering, Ningbo University, Ningbo, 315211 China*

[b] *Guangdong Key Laboratory of Intelligent Information Processing and Shenzhen Key Laboratory of Media Security, Shenzhen, 518060 China*

**e-mail: yandiqun@nbu.edu.cn*

Received February 27, 2018; revised May 17, 2018; accepted May 17, 2018

Abstract—Automatic speaker verification systems have achieved great performance and been widely adopted in many security applications. One of the important requirements for the verification system is its resilience to spoofing attacks, such as impersonation, replay, speech synthesis and voice conversion. Among these attacks, speech synthesis has a high risk to the verification systems. In this paper, a novel detection method for computer-generated speech, especially for HMM synthetic speech, is proposed. It is found that the wavelet coefficients in specified position show the obvious difference between the synthetic and natural speech. The logarithmic spectrum features are extracted from the wavelet coefficients and support vector machine is used as the classifier to evaluate the performance of our proposed algorithm. The experimental results over SAS corpus show that the proposed algorithm can achieve high detection accuracy and low equal error rate.

Keywords: speech synthesis, spoofing attack, wavelet transform, classification

DOI: 10.3103/S014641161901005X

1. INTRODUCTION

Automatic speaker verification [1] is the process of confirming or rejecting a speaker by his/her unique information. However, spoofing attacks [2], such as impersonation, replay, speech synthesis, and voice conversion, have caused a great threat to automatic speaker verification systems. Among these spoofing attacks, speech synthesis, also named text-to-speech (TTS), is a technique to generate artificial speech for the input text and it has been used widely in various speech applications. A speech synthesis system consists of two main parts: text analysis and speech waveform generation. There are two major approaches for speech waveform generation, including unit selection and hidden Markov model (HMM).

State-of-the-art HMM-based speech synthesis [3] can learn the models from relatively little speaker-specific data. A lot of past works [4, 5] have demonstrated the weakness of speaker verification systems to the HMM-synthesized voice. Since the dynamic changes of synthetic speech are usually less than those of natural speech, Satoh [6] took the intra-frame differences as the feature to detect HMM-based synthetic speech. In [7], Mel-cepstral coefficients (MFCC) are extracted to identify synthetic speech. The results show that the coefficients in high frequency bands of synthetic speech are smoothed when compared to natural speech. On the other hand, there are some schemes focusing on the differences between vocoders and natural speech. Generally, phase information is not considered in most of the synthetic vocoders. Hence, the differences in the phase spectra between natural and synthetic speech could be used to detect the synthetic process. These methods [8, 9] work well with the prior knowledge of the vocoders. In addition, considering the difficulty for modeling prosody during speech synthesis, the statistics of the fundamental frequency are used to detect the synthetic speech in [10, 11].

In this paper, a novel detection algorithm for HMM-based synthesis attack is proposed. Wavelet transform [12] is referred as “mathematical microscope,” which has been shown to have many advantages for speech signal analysis. From the experiment, we found that there are obvious differences in the detail wavelet coefficients between natural and synthetic speeches. Then the logarithmic spectrum features are

[1] The article is published in the original.

Table 1. Number of samples in the training and testing sets

Class	Natural Speech	Spoofing Speech	
		16 KHz	48 KHz
Training	725	1150	1116
Testing	349	574	557

Table 2. False acceptance rates for various speaker verification systems [13] (%)

Gender	Spoofing	GMM-UBM-5	GMM-UBM-50	JFA-5	JFA-50	PLDA-5	PLDA-50
Male	Baseline	4.05	0.09	2.76	1.25	1.41	1.16
	SS-LARGE-16	79.86	97.86	88.62	96.17	93.45	97.76
	SS-LARGE-48	97.35	99.95	97.62	98.93	99.12	99.09
Female	Baseline	11.10	0.66	6.24	2.47	1.52	0.99
	SS-LARGE-16	90.13	89.34	84.31	84.65	86.04	95.95
	SS-LARGE-48	98.52	99.28	90.58	94.28	94.80	98.35

extracted from the wavelet coefficients to detect the synthetic speech. The results show that the proposed algorithm can distinguish between synthetic and natural speech with high accuracy.

The rest of this paper is organized as follows. In Section 2, the corpus used in the experiment is described. The details of the proposed algorithm are provided in Section 3. In Section 4, we evaluate the suitability of the new feature set for detecting synthetic speech in various configurations. Finally, Section 5 concludes the paper.

2. SPOOFING CORPUS

The SAS corpus [13] is adopted in this work, which is published after the first ASV spoofing and countermeasures challenge [14]. This corpus contains natural speech and spoofed speech. The natural speech in the corpus is collected from 106 speakers (45 males, 61 females). For those spoofed speech generated by speech synthesis, an HMM-based synthesis algorithm [15] is adopted. In this work, two subsets of speech synthesis, called SS-LARGE-16K and SS-LARGE-48K are chosen to evaluate the performance of the proposed algorithm. The sampling rates of the two subsets are 16 kHz and 48 kHz, respectively. The summary of natural and spoofed speech for training and testing is shown in Table 1.

Table 2 shows the performance of various automatic speaker verification systems with two variants (-5 and -50 denote 5-utterance and 50-utterance enrollment scenarios) for synthetic spoofed speeches from SS-LARGE-16K and SS-LARGE-48K sets. JFA and PLDA in Table 2 represent Joint Factor Analysis and Probabilistic Linear Discriminant Analysis respectively. It can be seen that the false acceptance rates (FARs) for the baseline case which uses only natural speech (no spoofing attack) in Table 2 are less than 5% while the synthetic spoofing attack leads to FARs as high as 90%. The results show the weakness of current speaker verification systems and it is clear that countermeasures are needed.

3. DETECTION ALGORITHM FOR SYNTHETIC SPEECH

3.1. Wavelet-Domain Feature Extraction

As a time-frequency analysis tool, wavelet analysis is well applied in signal analysis and processing. It has made significant results in many fields, such as signal processing, pattern recognition, data compression, and so on. Since speech is a non-stationary signal and wavelet analysis has the characteristics of multi-resolution analysis, wavelet analysis is suitable for the speech signal.

In this work, discrete wavelet transform (DWT), which is the discrete version of the wavelet analysis, is adopted and can be defined by the following equation:

$$W(j,k) = \sum_j \sum_k x(n) \times 2^{-j/2} \psi\left(2^{-j} n - k\right), \quad (1)$$

Table 3. Coefficients of Haar, Daubechies4 and Symlet2 mother wavelets

Coefficients	Mother wavelet		
	Haar	Daubechies4	Symlet2
$l(0)$	−0.7071068	−0.1830127	−0.4829629
$l(1)$	0.7071068	−0.3169873	0.8365163
$l(2)$		1.1830127	−0.2241438
$l(3)$		−0.6830127	−0.1294095
$h(0)$	0.7071068	0.6830127	−0.1294095
$h(1)$	0.7071068	1.1830127	0.2241438
$h(2)$		0.3169873	0.8365163
$h(3)$		−0.1830127	0.4829629

where $x(n)$ is the input speech signal, $n, j, k \in Z$; j gives the dilation and k gives the translation. Here, ψ is a time function with finite energy and fast decay called the mother wavelet.

DWT analysis can be performed using a pyramidal algorithm related to multirate filterbanks. As a multirate filterbank, the DWT can be viewed as a constant Q filterbank with octave spacing between the centers of the filters. Each subband contains half the samples of the neighboring higher frequency subband. In the pyramidal algorithm, the signal is analyzed at different frequency bands with different resolution by decomposing the signal into a coarse approximation and detail information. The coarse approximation is then further decomposed using the same wavelet decomposition step. This is achieved by successive highpass and lowpass filtering of the time domain signal and is defined by the following equations:

$$cA_k = \sum_n x(n)l(2k - n), \tag{2}$$

$$cD_k = \sum_n x(n)h(2k - n), \tag{3}$$

where cA_k, cD_k are the outputs of the lowpass and highpass filters, l and h are impulse response coefficients of the mother wavelet. Table 3 shows the cases when the mother wavelets are Haar, Daubechies 4 and Symlet 2.

Let $S = \{s(i), 1 \le i \le L\}$ be a speech signal with L samples. According to the equation (2) and (3), 8-level DWT is applied to calculate the wavelet coefficients $cA_8, cD_8, cD_7, \ldots, cD_1$ as shown in Fig. 1, where cA_8 is the coarse signal of 8th-level and the detail signals of different levels are $cA_8, cD_8, cD_7, \ldots, cD_1$.

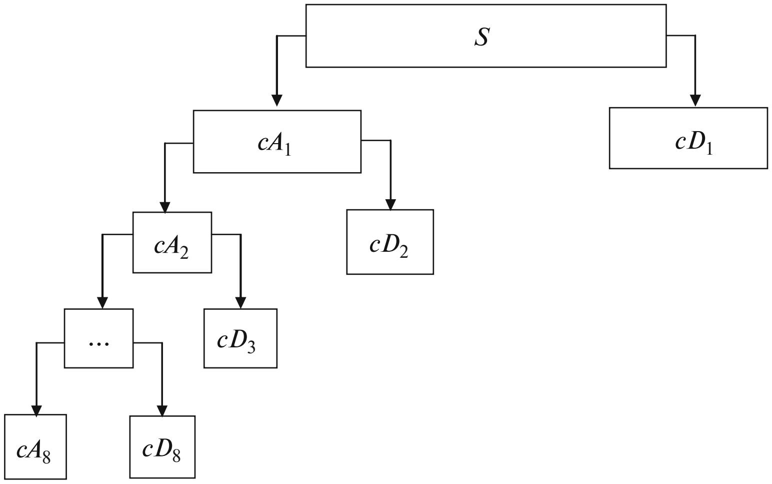

Fig. 1. 8-level wavelet decomposition.

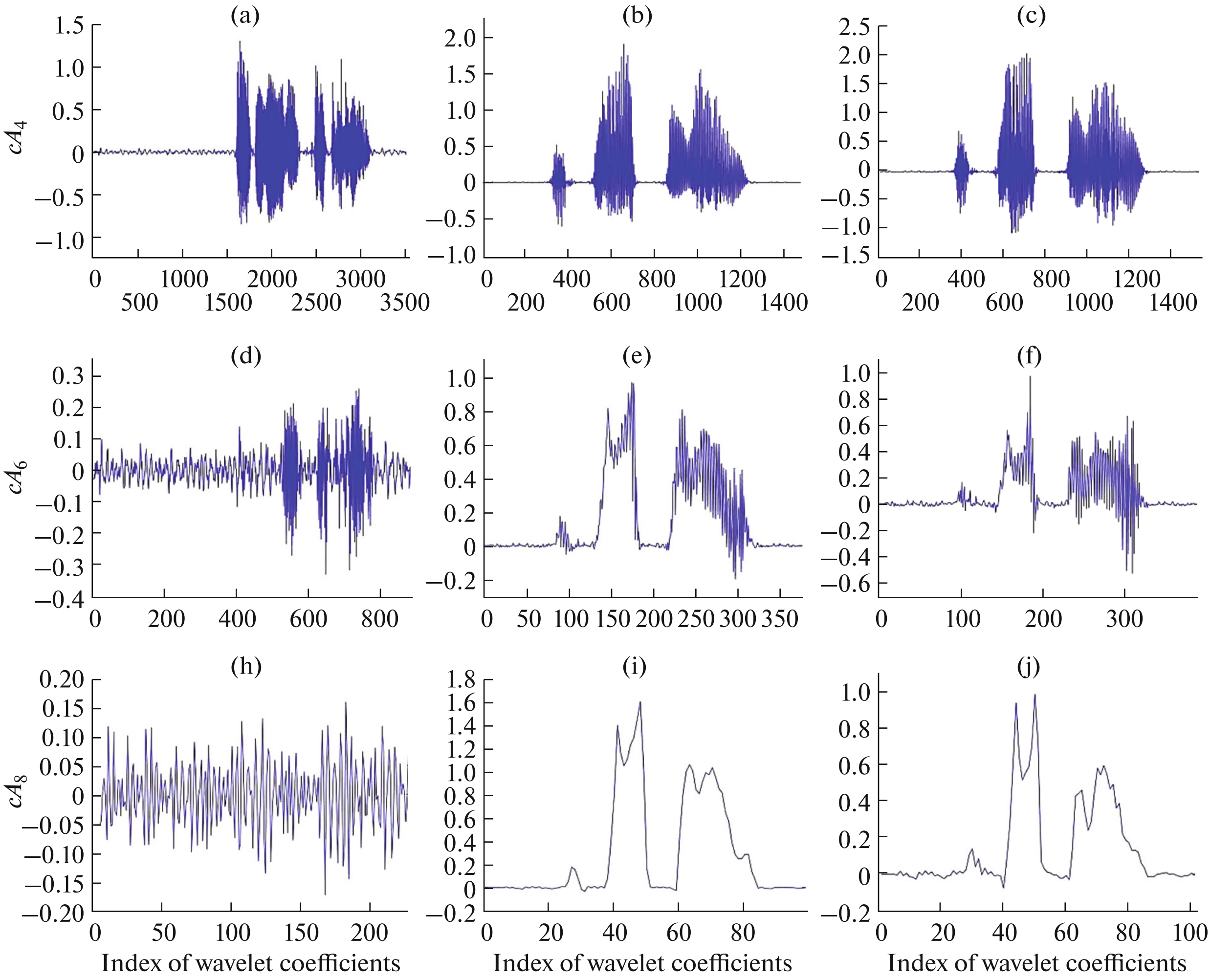

Fig. 2. Coarse wavelet coefficients. (a) cA_4 of Natural, 16 KHz; (b) cA_4 of Spoofing, 16 KHz; (c) cA_4 of Spoofing, 48 KHz; (d) cA_6 of Natural, 16 KHz; (e) cA_6 of Spoofing, 16 KHz; (f) cA_6 of Spoofing, 48 KHz; (h) cA_8 of Natural, 16 KHz; (i) cA_8 of Spoofing, 16 KHz; (j) cA_8 of Spoofing, 48 KHz.

The wavelet coefficients cA_4, cA_6, cA_8 of natural and synthetic speeches with 16 and 48 KHz are shown in Fig. 2. It can be seen that with the increase of the level of the wavelet decomposition, the difference between natural and synthetic speech becomes more obvious. It should be noted that for natural speech, the value of its coefficients fluctuates around zero, while most of the coefficients for synthetic speech is greater than zero. Additionally, in the case of the 8th level, it can be seen that the variance of the wavelet coefficients of natural speech is larger than synthetic speech. Hence, cA_8 is selected as the coefficients to extract the identification features.

To capture the difference between the natural and synthetic speech, the analysis in frequency domain is adopted. The frequency spectrum $F(k)$ of the cA_8 can be calculated by the Fourier Transform:

$$F(k) = \sum_{n=0}^{N-1} cA_8(n) e^{-j\frac{2\pi k}{N}n} = |F(k)| e^{i\phi(k)}, \ 0 \le k \le N-1, \tag{4}$$

where N is the length of Fourier Transform; $|F(k)|$ and $\phi(k)$ are the amplitude and phase spectrums respectively.

The spectrums of natural and synthetic speeches are shown in Fig. 3. It can be seen that the spectrum outlines of the two speeches have an obvious change. Compared with natural speech, synthetic speech has a big amplitude in low frequency and small amplitude in high frequency.

Then, the amplitude spectrum $|F(k)|$ is filtered by a filter-bank:

$$z_m = |F(k)| H(k)_m, \quad m \in \{1, 2, \dots, M\}, \tag{5}$$

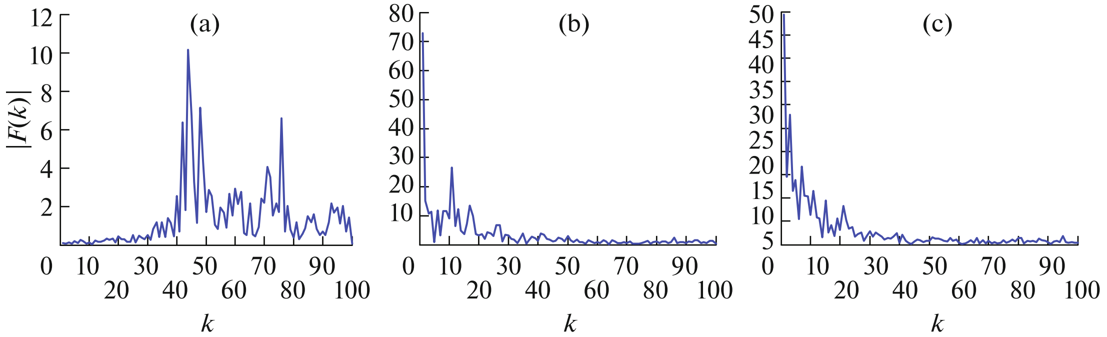

Fig. 3. Spectrums (a) Natural, 16 KHz (b) Spoofing, 16 KHz (c) Spoofing, 48 KHz.

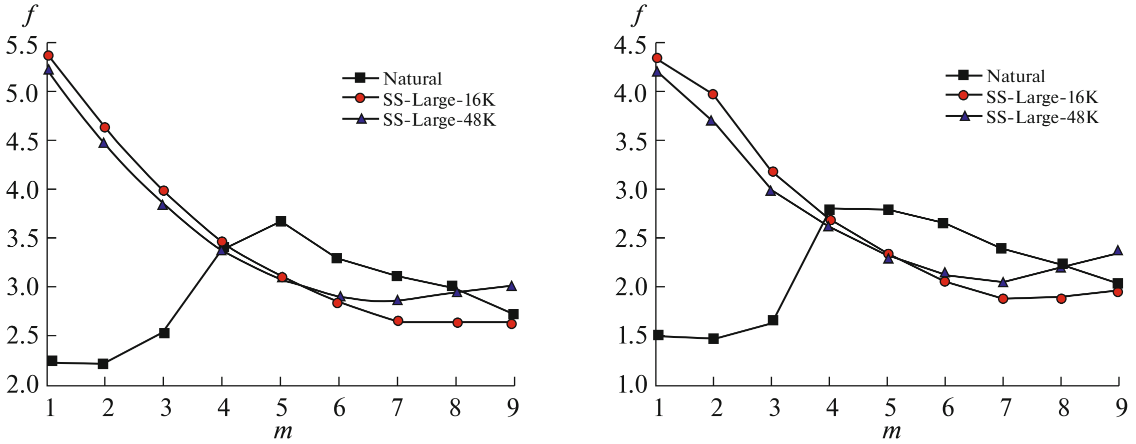

Fig. 4. Detection feature from natural and spoofed speech with various filters (a) Rectangle filter, (b) Triangle filter.

where M is the number of the filters in the filter bank, $H(k)_m$ denotes the transfer function of the m^{th} filter and z_m is the output of the filter. In this work, rectangular and triangular filters are used and their transfer functions are given by Eq. (6) and (7):

$$H_{\text{rect}}(k) = \begin{cases} 0 & k < g(m-1) \\ 1 & g(m-1) \le k \le g(m+1) \\ 0 & k > g(m+1) \end{cases}, \tag{6}$$

$$H_{\text{tri}}(k) = \begin{cases} 0 & k < g(m-1) \\ \dfrac{k - g(m-1)}{g(m) - g(m-1)} & g(m-1) \le k \le g(m) \\ \dfrac{g(m+1) - k}{g(m+1) - g(m)} & g(m) < k \le g(m+1) \\ 0 & k > g(m+1) \end{cases}, \tag{7}$$

where $g(m)$ is the center frequency of the mth filter and its definition is,

$$g(m) = \left\lceil \frac{N/2-1}{M+1} \right\rceil m, \quad m \in \{1, 2, \ldots, M\}. \tag{8}$$

Then the final features for detection f are calculated by mapping z to logarithmic domain by Eq. (9) and the dimension of the detection feature is M, which is the filter number of the filter bank. Figure 4 shows the mean value of f for 500 natural and spoofed speeches.

$$f_m = \ln(z_m). \tag{9}$$

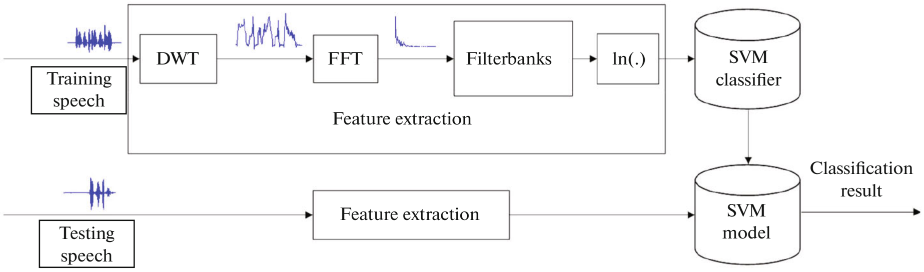

Fig. 5. Proposed detection algorithm.

3.2. Detection Algorithm for Synthetic Speech

The diagram of the proposed algorithm is shown in Fig. 5, which is based on the M -dimension features and support vector machine (SVM) classifier. The classification consists of training and testing stages. In the training, the dataset contains natural speech set is labeled as -1 and spoofed speech set is labeled as $+1$. The logarithmic spectrum features are extracted from the natural and spoofed speeches by Eqs. (1) to (9). The features are fed to SVM classifier. After 10-fold cross validation, a classification model is obtained. In the testing, for speeches in the testing set, the features are extracted and input them into the trained SVM. If the result of SVM classifier is -1, the testing speech is detected as a natural speech. Otherwise, the testing speech is a synthetic spoofed speech.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The corpus described in Section 2 is used in the experiment (see Table 1). In the experiment, Daubechies 4 (db4) is adopted as the wavelet basis function and the length of FFT is set as to 200. The filter number of filterbank M is set as to 9. The LIBSVM [16] is adopted as the SVM tool for training and testing. The SVM parameters are chosen as default for Radial Basis Function.

The results of accuracy (ACC) and equal error rate (EER) with rectangular and triangular filters are presented in Table 4. The definition of ACC is,

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100\%, \tag{10}$$

where TP is the occurrence that a synthetic speech is classified as synthetic one, TN is the occurrence that a natural speech is classified as natural one, FP is the occurrence that a natural speech is classified as a synthetic speech and FN is the occurrence that a synthetic speech is classified as a natural one. It can be seen that the average accuracy is higher than 99% when the sampling rates are 16 KHz and 48 KHz. The accuracy of [8] with the same experimental setup is 96.71%.

Meanwhile, the equal error rate (EER), which is a common measure to evaluate the algorithm's performance is also given in Table 4. It can be seen that all EER are less than 2%, which is better than the algorithm in [8]. Additionally, the EER for SS-Large-48 dataset and Rectangular Filterbank is 0.02%. It indicates that the proposed logarithmic spectrum features with rectangular filterbank can effectively distinguish the synthetic speech when high sampling rate is utilized during spoofing. The results indicate that the performance of the proposed algorithm is effective to distinguish the natural and synthetic speeches.

The detection performance is also evaluated by the receiver operating characteristics (ROC) curve, which reflects the results relating to the false alarm probability P_{FP} (false positives). Additionally, the area under ROC curve (AUC) is also used to measure the detection performance, which is defined as,

Table 4. Detection results of the proposed algorithm

Spoofing set	Rectangular Filterbank		Triangular Filterbank	
	ACC (%)	EER (%)	ACC (%)	EER (%)
SS-LARGE-16	98.98	1.30	99.16	1.12
SS-LARGE-48	99.63	0.02	98.88	1.30

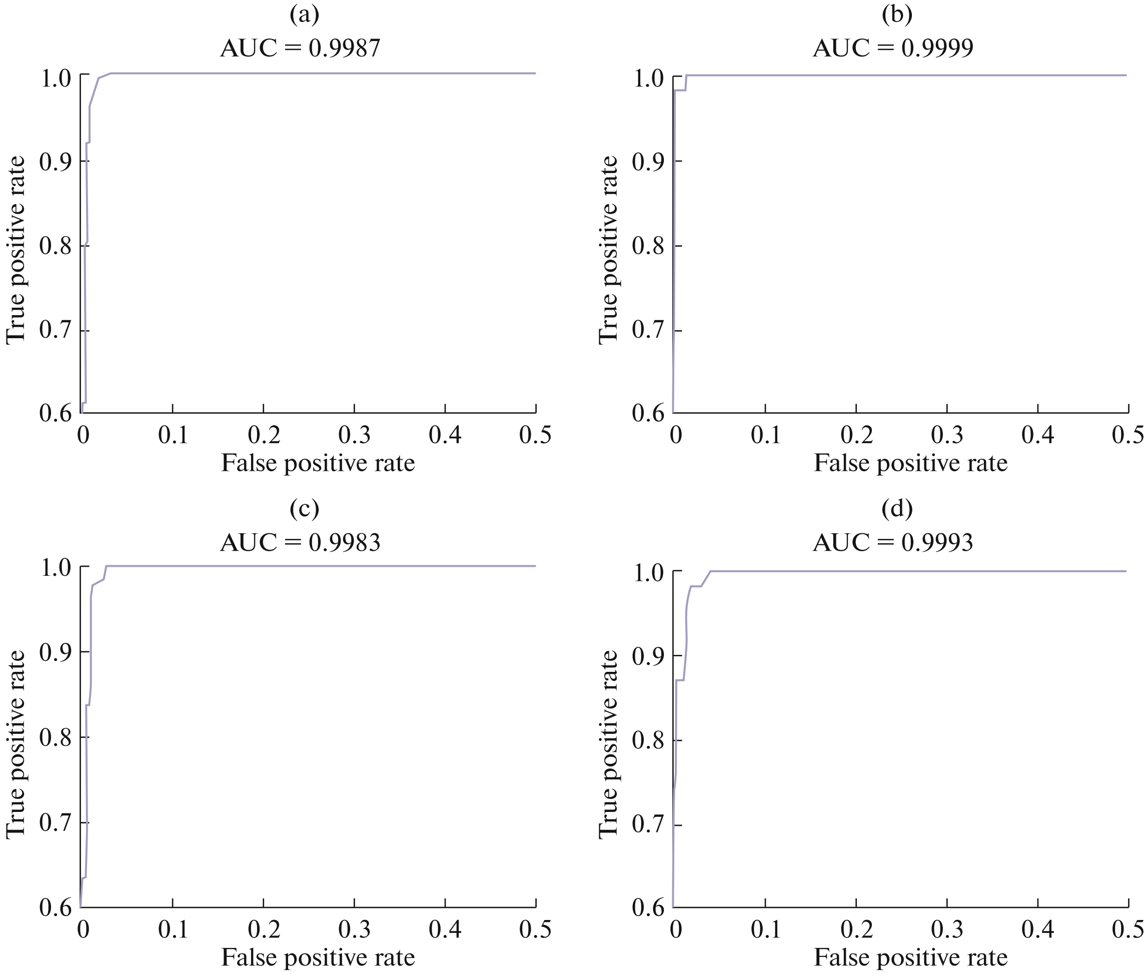

Fig. 6. ROC curves with various filters and sampling rates. (a) Rectangular filter, 16 KHz; (b) Triangular filter, 16 KHz; (c) Rectangular filter, 48 KHz; (d) Triangular filter, 48 KHz.

$$AUC = \int_0^1 P_{TP}(P_{FP})dP_{FP}, \quad (11)$$

where P_{TP} denotes the probability of true positives and P_{FP} denotes the probability of false positives.

Figure 6 shows the spoofed speech detection performance of the proposed algorithm in ROC curves. The values of AUC with rectangular and triangular filters are close to 1. It means that the proposed algorithm is successful in spoofed speech detection.

5. CONCLUSIONS

In this paper, an algorithm for identifying natural speech and synthetic speech is proposed. Logarithmic spectrum based on discrete wavelet transform is extracted as the detection features. The analysis of the features indicates that the low wavelet coefficients of the two kinds of speeches have an obvious difference. Therefore, the extracted features can be used to identify the natural and synthetic speeches. A detection algorithm based on SVM classifier is present in our work. Experimental results show that the proposed algorithm can achieve a high detection performance while keeping a low false alarm probability.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (grant nos. 61300055, U1736215, 61672302), Zhejiang Natural Science Foundation (grant nos. LY17F020010, LZ15F020002), Ningbo Natural Science Foundation (grant no. 2017A610123), Ningbo University Fund (grant no. XKXL1503) and K.C. Wong Magna Fund in Ningbo University.

REFERENCES

1. Kinnunen, T. and Li, H., An overview of text-independent speaker recognition: From features to supervectors, *Speech Commun.,* 2010, vol. 52, no. 2, pp. 12–40.
2. Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., and Li, H., Spoofing and countermeasures for speaker verification: A survey, *Speech Commun.,* 2015, vol. 66, pp. 130–153.
3. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J., Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm, *IEEE Trans. Audio Speech Lang. Process.,* 2009, vol. 17, no. 1, pp. 66–83.
4. Evans, N., Kinnunen, T., and Yamagishi, J., Spoofing and countermeasures for automatic speaker verification, *Proceedings of Annual Conference of the International Speech Communication Association,* 2013, pp. 925–929.
5. Alegre, F., Vipperla, R., Evans, N., and Fauve, B., On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals, *Proceedings of European Signal Processing Conference,* 2012, pp. 36–40.
6. Satoh, T., Masuko, T., Kobayashi, T., and Tokuda, K., A robust speaker verification system against imposture using an HMM-based speech synthesis system, *Proceedings of European Conference on Speech Communication and Technology,* 2001, pp. 759–762.
7. Chen, L.W., Guo, W.L., and Dai, R., Speaker verification against synthetic speech, *Proceedings of 7th International Symposium on Chinese Spoken Language Processing*, 2010, pp. 309–312.
8. De Leon, P.L., Pucher, M., Yamagishi, J., Hernaez, I., and Saratxaga, I., Evaluation of speaker verification security and detection of HMM-based synthetic speech, *IEEE Trans. Audio Speech Lang. Process.,* 2012, vol. 20, no. 8, pp. 2280–2290.
9. Wu, Z., Chng, E.S., and Li, H., Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition, *Proceedings of Annual Conference of the International Speech Communication Association,* 2012, pp. 1700–1703.
10. Ogihara, A., Unno, H., and Shiozakai, A., Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification, *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.,* 2005, vol. 88, no. 1, pp. 280–286.
11. De Leon, P.L., Stewart, B., and Yamagishi, J., Synthetic speech discrimination using pitch pattern statistics derived from image analysis, *Proceedings of Annual Conference of the International Speech Communication Association,* 2012, pp. 370–373.
12. Daubechies, I., The wavelet transform, time-frequency localization and signal analysis, *IEEE Trans. Inf. Theory,* 1990, vol. 36, no. 5, pp. 961–1005.
13. Wu, Z., De Leon, P.L., Demiroglu, C., and Khodabakhsh, A., Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance, *IEEE Trans. Audio Speech Lang. Process.,* 2016, vol. 24, no. 4, pp. 768–783.
14. Wu, Z., Khodabakhsh, A., Demiroglu, C., Yamagishi, J., Saito, D., Toda, T., and King, S., SAS: A speaker verification spoofing database containing diverse attacks, *Proceedings of International Conference on Acoustics, Speech and Signal Proceeding,* 2015, pp. 4440–4444.
15. Zen, H., Tokuda, K., and Black, A.W., Statistical parametric speech synthesis, *Speech Commun.,* 2009, vol. 51, no. 11, pp. 1039–1064.
16. Chang, C.C. and Lin, C.J., LIBSVM: A library for support vector machines, *ACM Trans. Intell. Syst. Technol.,* 2011, vol. 2, no. 3, pp. 1–27.