DETECTING SYNTHETIC SPEECH USING LONG TERM MAGNITUDE AND PHASE INFORMATION

Xiaohai Tian[1,2], *Steven Du*[1,3], *Xiong Xiao*[3], *Haihua Xu*[3], *Eng Siong Chng*[1,2,3] *and Haizhou Li*[1,4,5]

[1]School of Computer Engineering, Nanyang Technological University (NTU),Singapore
[2]Joint NTU-UBC Research Center of Excellence in Active Living for the Elderly, NTU, Singapore
[3]Temasek Laboratories, NTU, Singapore
[4]Human Language Technology Department, Institute for Infocomm Research, Singapore
[5]School of EE & Telecom, University of New South Wales, Australia

{xhtian, sjdu, xiaoxiong, haihuaxu, aseschng}@ntu.edu.sg, hli@i2r.a-star.edu.sg

ABSTRACT

Synthetic speech is speech signals generated by text-to-speech (TTS) and voice conversion (VC) techniques. They impose a threat to speaker verification (SV) systems as an attacker may make use of TTS or VC to synthesize a speakers voice to cheat the SV system. To address this challenge, we study the detection of synthetic speech using long term magnitude and phase information of speech. As most of the TTS and VC techniques make use of vocoders for speech analysis and synthesis, we focus on differentiating speech signals generated by vocoders from natural speech. Log magnitude spectrum and two phase-based features, including instantaneous frequency derivation and modified group delay, were studied in this work. We conducted experiments on the CMU-ARCTIC database using various speech features and a neural network classifier. During training, the synthetic speech detection is formulated as a 2-class classification problem and the neural network is trained to differentiate synthetic speech from natural speech. During testing, the posterior scores generated by the neural network is used for the detection of synthetic speech. The synthetic speech used in training and testing are generated by different types of vocoders and VC methods. Experimental results show that long term information up to 0.3s is important for synthetic speech detection. In addition, the high dimensional log magnitude spectrum features significantly outperforms the low dimensional MFCC features, showing that it is important to retain the detailed spectral information for detecting synthetic speech. Furthermore, the two phase-based features are found to perform well and complementary to the log magnitude spectrum features. The fusion of these features produces an equal error rate (EER) of 0.09%.

Index Terms— Spoofing attack, voice conversion, instantaneous frequency

1. INTRODUCTION

Speaker verification (SV) is the process of verifying the claimed identity of a user based on his/her speech signals. There are many potential applications of SV, such as access control for automatic services and phone banking [1]. Due to the high security requirement of these applications, SV system is expected to be very secure and should be able to perform even under malicious attacks. One common scenario of attacking an SV system is to cheat the SV system to get unauthorised access by simulating the speech signals of one of the user of the system. The speech signal of the user could be simulated by using various ways, such as voice conversion (VC) which convert one speakers voice to the target users speech, building a text-to-speech (TTS) system for the target user and generate speech using the TTS system, or recording the target users voice and reply it to the SV system. In this paper, we will focus on handling the first two ways, i.e. VC and TTS.

To address the attack by VC and TTS, one way is to improve the robustness of the SV system itself. In [2], the GMM-based SV system was used, and in [3], the joint factor analysis (JFA), GMM-UBM, VQ-UBM, GLDS-SVM, GMM-SVM and GMM-JFA systems have been used. However, one of the major issues on this kind of SV system is more focusing on the speaker identity verification but weak of synthetic speech detection.

Another way to address the VC and TTS attack to SV system is to add a screener that detect whether the incoming speech is natural speech or synthetic speech. If it is a natural speech, normal SV procedure will be carried out. If the incoming signal is detected as synthetic speech, the system will directly reject it hence protect the SV system from attack. In this work, we focus on the problem of synthetic speech detection. Several types of synthetic spoofing speech have been studied in the past. The synthetic speech from Hidden Markov Model (HMM) based TTS system was studied in [4, 5, 6] and adapted statistical speech synthesis system in [7]. The synthetic speech generated by VC techniques [8, 9, 10], have been studied in [11, 12, 3, 13, 14, 15]. Nevertheless, most of previous works heavily rely on GMM-based clustering, which can only handle low-dimensional feature without context information. The limited input information will reduce the system performance.

An important topic of spoofing speech detection is the selection of proper features that is able to differentiate natural speech from synthetic speech. Mel-frequency cepstrum coefficient (MFCC), which is the most commonly used acoustic feature for representing the short-term power spectrum, is wildly used in previous works [2, 3, 4, 5, 6]. Motivated by the fact that VC or TTS systems may causes artefacts to the phase spectrum of speech, cosine-normalized phase and modified group delay (MGD) phase features were used in [16, 17]. Moreover, intending to detect the temporal distortion of synthetic speech, modulation feature was introduced in [18]. Most of the features used in previous works are

This research is supported in part by Interactive and Digital Media Programme Office (IDMPO), National Research Foundation (NRF) hosted at Media Development Authority (MDA) of Singapore (Grant No.: MDA/IDM/2012/8/8-2 VOL 01).

low-dimensional features, which only model the formant shape of speech (MFCC) and coarse shapes of phase (MGD).

In this paper, we focus on feature study for synthetic speech detection. We argue that the low-dimensional features does not carry the detailed information which tells the difference between natural and synthetic speech. This is because most VC and TTS methods focused on reducing the artefacts on spectrum shapes which is represented by low-dimensional cepstral coefficients. Hence, it is necessary to look at high resolution representation of speech that may reveal the artefacts of VC and TTS systems. In this study, we directly use the full details of log magnitude spectrum as features for detecting synthetic speech. We also use two phase-derived features that represent the high resolution phase information of speech. Besides using high resolution information, we also include long term temporal information by using a sequence of frames rather than a single frame as the input of the synthetic speech detection system. It is observed that up to 0.3s of input window is necessary for optimal detection performance. To handle the high dimensional input which is the result of using high resolution and long temporal window, we use neural networks (NN) as the natural/synthetic speech classifier. Unlike conventional classifier used in speech processing, such as Gaussian mixture model (GMM), NN does not have limitation in input dimension.

The paper is organized as follows. The various features used in the study are introduced in section 2. In section 3, neural network classifier is briefly introduced. The experimental settings and objective evaluation are presented in section 4. The paper will be concluded in section 5.

2. FEATURE EXTRACTION

In this study, 5 types of features are used, including two magnitude based features, i.e. high-dimensional log magnitude spectrum and low-dimensional MFCC, and 3 phase based features, including high-dimensional instantaneous frequency (IF), MGD, and low-dimensional cepstral features derived from MGD. In this section, we will briefly describe the extraction of these features.

2.1. Magnitude spectrum

The log magnitude spectrum of speech signal could be obtained by applying short-time Fourier transform (STFT). A speech signal is divided into 25ms long overlapping data frames, DC offset removed, Hamming windowed, and then applied FFT. Given a speech signal $x(n)$, the complex spectrum can be expressed as:

$$\mathrm{X}(t,\omega) = |\mathrm{X}(t,\omega)|e^{j\theta(t,\omega)}, \tag{1}$$

where, $|\mathrm{X}(t,\omega)|$ and $\theta(t,\omega)$ are the magnitude and phase spectrum at frame t and frequency ω, respectively. The overlap between two adjacent frames is set to 15ms. The log magnitude spectrum feature vectors is defined to be $\mathbf{x} = [log(|\mathrm{X}(t,0)|),\ldots,log(|\mathrm{X}(t,\pi)|)]^\top$ As we will experiment with speech data sampled at 16kHz, the FFT length is chosen to be 512, and each feature vector will have 257 dimensions.

2.2. Mel-frequency cepstrum coefficient (MFCC)

The MFCC features [19] can be seen as a compact representation of the log magnitude spectrum. First, Mel-frequency scaled filter banks are computed by summing neighbouring frequency bins of magnitude spectrum together according to the nonlinear Mel scale, resulting in 23 filter bank coefficients. Then logarithm is applied and discrete cosine transform is applied to reduce the dimensionality further down to 13. Finally, the delta and acceleration of features which carries temporal information up to 0.1s are appended to the 13 MFCC features to form the 39 feature vectors. MFCC features are widely used in ASR/SV systems and some previous synthetic speech detection systems [2, 3, 14]. They will be used in our study as baseline features.

2.3. Instantaneous frequency derivative (IFD)

In order to test the phase continuity on time domain, the instantaneous frequency derivative (IFD) [20] feature is used. The IFD feature extraction is based on the phase spectrum of the speech signal $\theta(t,\omega)$ and defined as the derivatives of the phase spectrum w.r.t. time:

$$\begin{aligned} IFD(t,\omega) &= \frac{1}{2\pi}\frac{d\theta(t,\omega)}{dt} \\ &\approx \frac{1}{2\pi}(\theta(t,\omega) - \theta(t-1,\omega)). \end{aligned} \tag{2}$$

Each element of IFD is added an integer multiples of π to make its value within the interval $[-\pi, \pi]$. Unlike the original phase spectrum that hardly show any patterns, there is clear patterns in the IFD, making it possible to use them as features for synthetic speech detection.

2.4. Modified group delay (MGD)

Another phase feature named modified group delay (MGD) [20] is also used in this work to detect the non-linearity of the phase spectrum. For the speech signal $x(n)$, MGD feature can be calculated as follows (the frame subscripts are ignored for clearer formulas):

1) Compute the STFT of $x(n)$ and $nx(n)$ separately, denoted as $X(\omega)$ and $Y(\omega)$
2) Compute the smoothed spectrum $|X(\omega)|$, denoted as $S(\omega)$
3) Compute the MGD as:

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\gamma}} \tag{3}$$

4) Then, we reshape it as:

$$\hat{\tau}(\omega) = \frac{\tau(\omega)}{|\tau(\omega)|}|\tau(\omega)|^{\alpha} \tag{4}$$

Based on our experimental results, we set the γ and α as 1.2 and 0.4 respectively.

2.5. Modified group delay cepstral coefficient (MGD-Cep)

The MGD-Cep extraction is based on the MGD feature $\hat{\tau}(\omega)$. The processing is similar to the extraction of MFCC features from magnitude spectrum and the main difference is that the MGD is used to in place of the magnitude spectrum. We first compute the filterbank energies (FBE) by apple a Mel-frequency filter bank to $\hat{\tau}(\omega)$. Then, the MGD-Cep could be obtained by applying DCT to FBE. The MGD-Cep features have been used in [16] for synthetic speech detection.

3. NEURALNETWORK-BASEDNATURAL/SYNTHETIC SPEECH CLASSIFIER

As high dimensional feature vectors are used and many feature vectors are concatenated to form the input of the synthetic speech detection system, the systems input dimensions are typically very high, e.g. up to 10,000. Such high dimensional features cannot be modelled by conventional classifiers in speech processing, such as GMM-based generative models. Instead, we choose neural networks (NN) to estimate the posterior probabilities of the speech/synthetic speech classes given the input.

Although the synthetic speech detection is a detection problem, we treat it as a two class classification problem during NN training. Given an input sample, which is usually a concatenation of many frames of feature vectors, the NN predict the posterior probabilities of the input being generated from natural speech. As this is a two class problem, the two posteriors will sum to 1, so it is sufficient to just retrain the posterior of natural speech. During testing, the posterior probabilities of natural speech will be used as the score for synthetic speech detection. Intuitively, a natural speech will usually have high scores and synthetic speech will usually have low scores.

For each test utterance, for each frame of 25ms, a patch of feature vectors are extracted by concatenating the feature vectors around the current frame. A score is generated for every frame, and the final score for the utterance is the mean of the frame level scores. A simple VAD is used to discarded the scores from silence frames. Finally, one score is generated for each test utterance and the equal error rate (EER) can be obtained by selecting a proper threshold and the detection error tradeoff (DET) curve could be plotted.

4. EXPERIMENTAL SETUP

4.1. Corpus

Four speakers with US accent of CMU-ARCTIC database [21], (*bdl*, *rms*, *slt* and *clb*) were used in this work. The speech signals were sampled as 16 kHz. Each of the speaker has about 1000 utterances, and the label of these utterances are the same across all speakers.

The synthetic speech is generated by applying the following four techniques on the natural speech in CMU-ARCTIC.

1. **AHOcoder-syn**: The AHOcoder [22] was used for speech analysis and reconstruction, without feature transformation.
2. **STRAIGHT-syn**: The STRAIGHT [23] was used for speech analysis and reconstruction, without feature transformation.
3. **GMM-VC**: The JD-GMM with maximum likelihood parameter generation method as proposed in [9]. MCC features were used to train the model, the optimal number of Gaussian mixtures was 64.
4. **CFW-VC**: The weighted correlation-based frequency warping [24] with GMM-based residual compensation. Spectral envelopes were used to find the warping function, based on formant segmentation. The segment boundary shift was constrained within 100 Hz. Only voiced frames were transformed in this method.

For the two VC methods, STRAIGHT was also used for feature extraction. 513-dimensional spectrum and log F_0 were extracted by STRAIGHT. 25-dimensional Mel-Cepstral Coefficients (MCCs) and 15-dimensional linear spectrum frequencies (LSFs) were also used for the spectrum. In all the conversion methods, we used the same frame alignment, which was obtained by performing DTW on the MCC feature sequence.

To create an open test scenario, we use the speech of *bdl* and *clb* for training, and speech of *slt* and *rms* for testing. In addition, only synthetic speech generated by STRAIGHT-syn and GMM-VC are used during training, hence AHOcoder-syn and CFW-VC can serve as unseen synthetic speech types during test.

4.2. Baseline method

The GMM-based synthetic speech detection system [16] was used as a baseline in our work. The log likelihood ratio of natural and synthetic speech classes are used for the detection problem:

$$\boldsymbol{\Gamma}(O) = \log p(O|\lambda_{synthetic}) - \log p(O|\lambda_{nature}), \qquad (5)$$

where, O is the observation feature, $\lambda_{synthetic}$ and λ_{nature} are the GMM model of synthetic and nature speech respectively. The number of Gaussian component is set as 1024.

5. EVALUATION AND DISCUSSION

As the synthetic speech detection is a detection problem, we use two related evaluation metrics, i.e. the EER value and DET curve. The EER is obtained by selecting a specific threshold (an operating point of the detection system) such that the miss rate and false alarm rate are equal. On the other hand, the DET curve plot the miss rate against the false alarm rate at all possible operating points. While EER gives a simple indicator of whether a system performs well, the DET curve presents a more complete picture of the system performance.

5.1. The effect of input context size

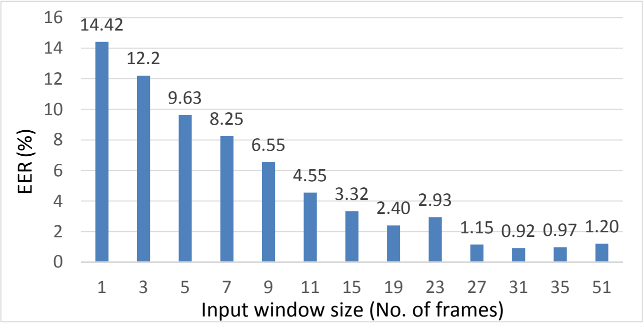

Fig. 1. Equal error rate of log spectrum as a function of the input context size.

The effect of using long term temporal information in NN based systems is examined. Fig. 1 presents the EER results using different context size with log magnitude spectrum. It is observed that the EER gradually reduces as the context size increases from 1 to 31 frames, which amounts to 310ms of speech information. The performance saturates at about 31 frames and longer context such as 51 frames even degrades the performance slightly which is perhaps due to overfitting. The best EER=0.92% is obtained at 31 frames context size. Besides log magnitude spectrum, we also found that the temporal information is also important for synthetic speech detection when other types of features are used, such as IFD and MGD. These observations show that long term temporal information is useful for detecting synthetic speech.

5.2. Comparison of different features

Table 1. *Equal error rate (EER, %) of detection performance of different systems.*

No.	Systems	Natural against individual synthetic data				Natural
		STRAIGHT-syn	GMM-VC	AHOcoder-syn	CFW-VC	against
Low dimensional features						
1	MFCC(1)-GMM	2.65	0.35	51.52	4.33	21.34
2	MFCC(1)-NN	17.51	0.29	51.57	33.28	30.85
3	MFCC(31)-NN	4.44	2.74	50.77	12.15	22.41
4	MGD-Cep(31)-NN	22.99	1.21	33.68	7.93	19.48
High dimensional features						
5	Magnitude(31)-NN	**0.09**	**0.00**	0.22	2.14	0.92
6	MGD(31)-NN	9.14	0.31	5.17	5.04	5.78
7	IF(31)-NN	0.77	0.04	**0.13**	**0.80**	0.54
Fusion						
	Fusion(5+6)	0.09	0.00	0.00	0.53	0.23
	Fusion(5+7)	0.04	0.00	0.04	0.35	0.13
	Fusion(6+7)	1.02	0.04	0.15	0.71	0.68
8	Fusion(5+6+7)	**0.04**	**0.00**	**0.00**	**0.18**	**0.09**

The EER obtained by all features are shown in Table 1. The systems 1-4 uses low dimensional features, while the systems 5-7 uses high dimensional features. We also presented the different ways of fusing the systems 5-7 and system 8 is the best fusion results.

From the table, a general observation is that the first 4 systems do not performs well and their average EER are about 20-30%. This is true for both NN and GMM-based systems. It is also true when both 1 frame and 31 frames of input features are used. The results clearly show that low dimensional features are not effective in detecting synthetic speech.

The systems using high-dimensional features, namely log magnitude spectrum (system 5), MGD (system 6), and IF (system 7), all performs much better than the systems using low dimensional features. Comparing system 4 and 6, the only difference between them is that the MGD-Cep used in system 4 is a smoothed and dimension reduced version of MGD used in system 6. However, the EER of system 6 (5.78%) is much lower than that of system 4 (19.48%). Although more parameters are used in system 6 due to the much higher input dimension, we believe that the real factor that caused the different performance between system 4 and 6 is that high dimensional MGD which contains detailed information of the input speech is used in system 6. Similarly, system 5 and 3 are similar to each other, the main difference between them is that system 5 uses the full detailed information of the magnitude spectrum while system 3 only uses the formant shape of the magnitude spectrum that is represented by the MFCC features. Again, the system with detailed information performs much better. Hence, we can conclude that the detailed magnitude or phase information of speech signal is the key for achieving good synthetic speech detection performance.

The fusion of system 5-7 are shown in the last 4 rows of Table 1. The results show that although system 6 performs much worse than system 5 and 7, it provides complementary information and the best fusion results are obtained by fusing all systems. The fusion is performed by averaging the scores of individual systems.

The detection error tradeoff (DET) curves of our eight systems are presented in Figure 2. First, the curves of second system (MFCC(1)-NN) and third system (MFCC(31)-NN) show that the long term imformation helps to imporve the system performance. Secondly, the systems with high-dimensional features (the system 5-7) performe much better than the systems using low-dimensional features (the system 1-4). Moreover, the system performance is further improved with the combination of three high-dimensional features with long term information (system 8).

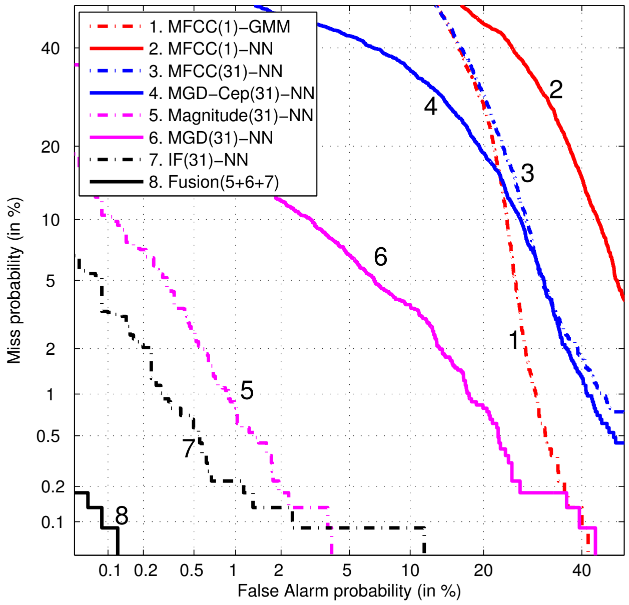

Fig. 2. DET curve of synthetic detection performance of different systems.

6. CONCLUSION

In this paper, we have conducted two studies for synthetic speech detection problem, including the use of long term temporal information and high dimensional feature vectors that contain detailed magnitude and phase information of the input speech signal. Results show that both long term temporal information and detailed speech information are vital for accurate detecting of synthetic speech. In addition, the log magnitude spectrum feature is shown to be complementary to the phase derived features, such as MGD and IF.

Currently, we have achieved high performance in detecting synthetic speech using microphone speech in clean environment. However, many real applications may involve noisy environments and speech coding/transmission, which introduces distortion even to natural speech. In the future, we will investigate the robust detection of synthetic speech in these scenarios.

7. REFERENCES

[1] Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A Reynolds, "A tutorial on text-independent speaker verification," *EURASIP journal on applied signal processing*, vol. 2004, pp. 430–451, 2004.

[2] Jean-François Bonastre, Driss Matrouf, and Corinne Fredouille, "Artificial impostor voice transformation effects on false acceptance rates.," in *Proc. INTERSPEECH*, 2007, pp. 2053–2056.

[3] Tomi Kinnunen, Zhi-Zheng Wu, Kong Aik Lee, Filip Sedlak, Eng Siong Chng, and Haizhou Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4401–4404.

[4] Takashi Masuko, Takafumi Hitotsumatsu, Keiichi Tokuda, and Takao Kobayashi, "On the security of hmm-based speaker verification systems against imposture using synthetic speech," in *Proc. Eurospeech*, 1999.

[5] Takashi Masuko, Keiichi Tokuda, and Takao Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch.," in *Proc. INTERSPEECH*, 2000, pp. 302–305.

[6] Takayuki Satoh, Takashi Masuko, Takao Kobayashi, and Keiichi Tokuda, "A robust speaker verification system against imposture using an hmm-based speech synthesis system," in *Proc. EUROSPEECH*, 2001, pp. 759–762.

[7] Phillip L De Leon, Michael Pucher, and Junichi Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," 2010.

[8] Yannis Stylianou, Olivier Cappé, and Eric Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[9] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[10] Daniel Erro, Asunción Moreno, and Antonio Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[11] Qin Jin, Arthur R Toth, Alan W Black, and Tanja Schultz, "Is voice transformation a threat to speaker identification?," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 4845–4848.

[12] Qin Jin, Arthur R Toth, Tanja Schultz, and Alan W Black, "Voice convergin: Speaker de-identification by voice transformation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 3909–3912.

[13] Zhizheng Wu and Haizhou Li, "Voice conversion and spoofing attack on speaker verification systems," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2013, pp. 1–9.

[14] Zhizheng Wu, Anthony Larcher, Kong-Aik Lee, Engsiong Chng, Tomi Kinnunen, and Haizhou Li, "Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints.," in *Proc. INTERSPEECH*, 2013, pp. 950–954.

[15] Zhizheng Wu and Haizhou Li, "Voice conversion and spoofing attack on speaker verification systems," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2013, pp. 1–9.

[16] Zhizheng Wu, Chng Eng Siong, and Haizhou Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. INTERSPEECH*, 2012.

[17] Zhizheng Wu, Tomi Kinnunen, Eng Siong Chng, Haizhou Li, and Eliathamby Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2012, pp. 1–5.

[18] Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li, "Synthetic speech detection using temporal modulation feature," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7234–7238.

[19] Steven Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[20] Leigh D Alsteris and Kuldip K Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.

[21] John Kominek and Alan W Black, "The CMU Arctic speech databases," in *Proc. ISCA Workshop on Speech Synthesis*, 2004.

[22] Daniel Erro, Iñaki Sainz, Eva Navas, and Inma Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, 2014.

[23] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[24] Xiaohai Tian, Zhizheng Wu, Siu Wa Lee, and Eng Siong Chng, "Correlation-based frequency warping for voice conversion," in *Proc. 9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2014, pp. 211–215.