Spoofing Speech Detection Using High Dimensional Magnitude and Phase Features: the NTU Approach for ASVspoof 2015 Challenge

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

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

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

Abstract

Recent improvement in text-to-speech (TTS) and voice conversion (VC) techniques presents a threat to automatic speaker verification (ASV) systems. An attacker can use the TTS or VC systems to impersonate a target speaker's voice. To overcome such a challenge, we study the detection of such synthetic speech (called spoofing speech) in this paper. We propose to use high dimensional magnitude and phase based features and long term temporal information for the task. In total, 2 types of magnitude based features and 5 types of phase based features are used. For each feature type, we build a component system using a multilayer perceptron to predict the posterior probabilities of the input features extracted from spoofing speech. The probabilities of all component systems are averaged to produce the score for final decision. When tested on the ASVspoof 2015 benchmarking task, an equal error rate (EER) of 0.29% is obtained for known spoofing types, which demonstrates the highly effectiveness of the 7 features used. For unknown spoofing types, the EER is much higher at 5.23%, suggesting that future research should be focused on improving the generalization of the techniques.

Index Terms: Spoofing attack, voice conversion, automatic speaker verification, phase feature, ASVspoof 2015.

1. Introduction

Automatic speaker verification (ASV) is the verification of a speaker's identify based on his/her speech signals [1]. There are many applications of ASV technology, such as access control. Due to the high security requirement of these applications, ASV system is required to be robust against malicious attacks.

Recently, advancement in text-to-speech (TTS) and voice conversion (VC) technologies makes possible high quality synthesis of any speaker's voice, provided that certain amount of training data of the speaker is available. Such capability imposes a significant threat to ASV systems, as the attacker can compromise the ASV system by using TTS or VC systems to synthesize the speech of the target speaker. To address this threat, the ASVspoof 2015 challenge [2] is introduced as a benchmark to measure the progress in spoofing speech detection. This paper describes the Nanyang Technological University (NTU) team's effort in participation of the open challenge.

There are two major ways to address the threat of spoofing attack [3], one is to improve the robustness of the ASV system itself. In [4, 5], various ASV systems are studied in terms of their robustness against spoofing attacks, such as the joint factor analysis (JFA), Gaussian mixture model-universal background model (GMM-UBM), etc. These works are more focused on the speaker identity verification but less on the spoofing speech detection. Another way to address the threat is to add a screener that detects whether the incoming speech is natural speech or synthetic speech. If a synthetic speech is detected, the screener will directly reject it, hence protecting the ASV facility.

Several spoofing speech detection methods have been proposed in the past. The synthetic speech from hidden Markov model (HMM) based TTS system was studied in [6–8] and speaker adapted statistical TTS system was studied in [9]. The synthetic speech generated by VC techniques [10–12], have been studied in [5, 13–16]. Most of these works heavily rely on GMM-based classifier, which only allows the use of low dimensional features for spoofing speech detection, such as Mel-frequency cepstrum coefficient (MFCC) [4–8] and modified group delay based features [17–19].

In this paper, we propose to use high dimensional speech features derived from both magnitude and phase spectra for detecting spoofing speech. In addition, we concatenate feature vectors within a window to incorporate long term temporal information. To handle the high dimensional feature vectors, multilayer perceptron (MLP) neural network is used to predict the posterior probabilities of the test speech being spoofing speech.

The rest of the paper is organized as follows. The ASVspoof 2015 challenge is briefly introduced in section 2. The overview of our system is proposed in section 3. The extraction of various features are described in section 4, followed by experimental results and discussions in section 5. Finally, conclusion and future works are presented in section 6.

2. Task Description

The ASVspoof 2015 challenge is designed to be a benchmarking task for spoofing speech detection. Three data sets are provided, including training, development, and evaluation sets, and their statistics are listed in Table 1. There are totally 10 spoofing types, from S1 to S10. The first 5 types (S1-S5) exist in all data sets and are called "known types" as they are used for system building. The last 5 types (S6-S10) are "unknown types" only exist in the evaluation set. The known types are: S1) frame selection based VC; S2) VC that modifies the first Mel cepstral coefficient; S3) HMM-based TTS adapted to target speaker using 20 sentences; S4) same as S3 but using 40 adaptation sentences

Table 1: Statisitics of the ASVspoof 2015 data sets. The speakers in the three data sets are non-overlapping.

Set	#Speaker		#Utterances		Spoofing Types
	M	F	Genuine	Spoofed	
Train	10	15	3,750	12,625	S1-S5
Dev	15	20	3,497	49,875	S1-S5
Eval	20	26	9,404	184,001	S1-S10

per speaker; S5) GMM-based VC considering global variance.

As this is a detection problem, the official system performance measure of ASVspoof 2015 is equal error rate (EER), which is the rate when the false alarm rate is equal to the miss rate. To obtain a more complete view of system performance, we also use the detection error tradeoff (DET) curve [20] for system evaluation on the development data for which we know the ground truth of whether a sentence is natural or spoofing. For more details of the task, the readers are refer to [2].

3. The NTU Approach

The proposed system is illustrated in Fig. 1. There are multiple component systems, each using a specific type of features. The scores of component systems are fused to generate the final score, one number for each test utterance. The final score is used to decide whether a test utterance is spoofing speech.

The use of multiple component systems is motivated by the fact that no single type of feature is able to detect all types of spoofing speech. As the spoofing speech may be generated by various types of TTS and VC methods, they may have different artefacts. A single type of features is usually good at detecting certain types of artefacts, but not all. Hence, it is wise to build multiple "expert" systems, each focusing on detecting certain types of artefacts by using one type of features, and to fuse their scores. The fusion can take many forms, such as voting or linear combination. In this study, we compute the final score as the simple average of component systems' scores instead of weighted average to avoid overfitting to the development data which does not contain the "unknown" spoofing types.

For each component system, an MLP is trained to predict the posterior probability of the input feature patch extracted from spoofing speech. A feature patch consists of 51 consecutive feature frames. We use 0.025s frame length and 0.01s frame shift, hence a feature patch covers about 0.5s temporal context. From each utterance, a sequence of feature patches are extracted with 50 frames overlap. Due to the use of long temporal context, the final feature vector (obtained by converting a feature patch into vector form) is of very high dimension, e.g. more than 10,000 elements. Such high dimensional features cannot be modelled by conventional classifiers in speech processing, such as GMM-based generative models. Therefore, we use MLP as the classifiers since it has no limitation on input feature dimensions. During the training, the spoofing speech detection is treated as a 2-class classification problem. The MLP contains 1 hidden layer with 3,072 sigmoid nodes and is trained to predict whether a feature patch is from natural or spoofing speech. During testing, the posterior probabilities of all feature patches of a test speech are averaged to produce one single posterior for each component system. A pitch based voice activity detector (VAD) is used to discard scores of silence patches. Finally, the component system scores are averaged to produce the final score for spoofing detection.

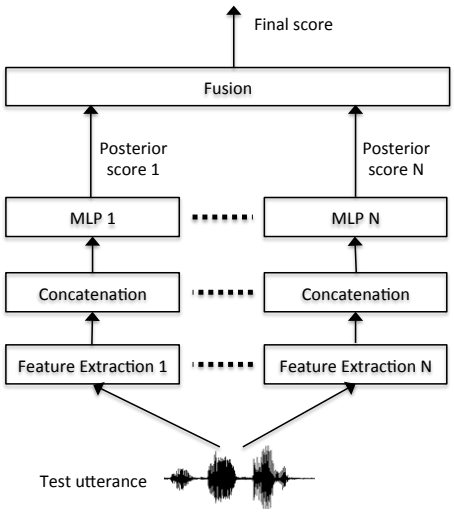

Figure 1: System architecture

4. Feature Extraction

In this study, 7 types of features are used, including 2 magnitude based features and 5 phase based features. These features are described in detail in the following sections and their examples are shown in Fig. 2. All features are extracted from the short time Fourier transform of the speech signal, which can be expressed as:

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

where, $|\mathrm{X}(t,\omega)|$ and $\theta(t,\omega)$ are the magnitude and phase spectra at frame t and frequency bin ω, respectively. The frame length and frame shift are set to 0.025s and 0.01s, respectively, except for the pitch synchronous phase (PSP) features for which variable frame lengths and shifts are used. As the speech signal of ASVspoof 2015 challenge is sampled at 16kHz, a frame length of 0.025s contains 400 samples, so the FFT length is set to 512. By retraining only half of the symmetric spectrum, the dimensionality of both phase and magnitude based features will be 256. After concatenating 51 frames of features, the input dimension of all component system MLPs is $51 \times 256 = 13,056$. In the following sections, the details of extracting the 7 types of featuers will be described.

4.1. Log Magnitude Spectrum (LMS)

The log magnitude spectrum feature is simply $\mathrm{LMS}(t,\omega) = \log(|\mathrm{X}(t,\omega)|)$. An example of LMS feature is shown in Fig. 2a. The magnidue spectrum contains all the detailed information about the speech signal, such as formant, pitch, and harmonic structure of vowel sounds. The logarithm is used to reduce the dynamic range of the magnitude spectrum, making them suitable to use as features.

4.2. Residual Log Magnitude Spectrum (RLMS)

The formant information in LMS is important for speech recognition, but may not be useful for spoofing detection as most of the spoofing techniques, such as VC or TTS, are good at modelling the formant of speakers. To remove the effect of formant, we also extract the LMS using (1) from the residual waveform of linear predictive coding (LPC). In the LPC analysis of speech

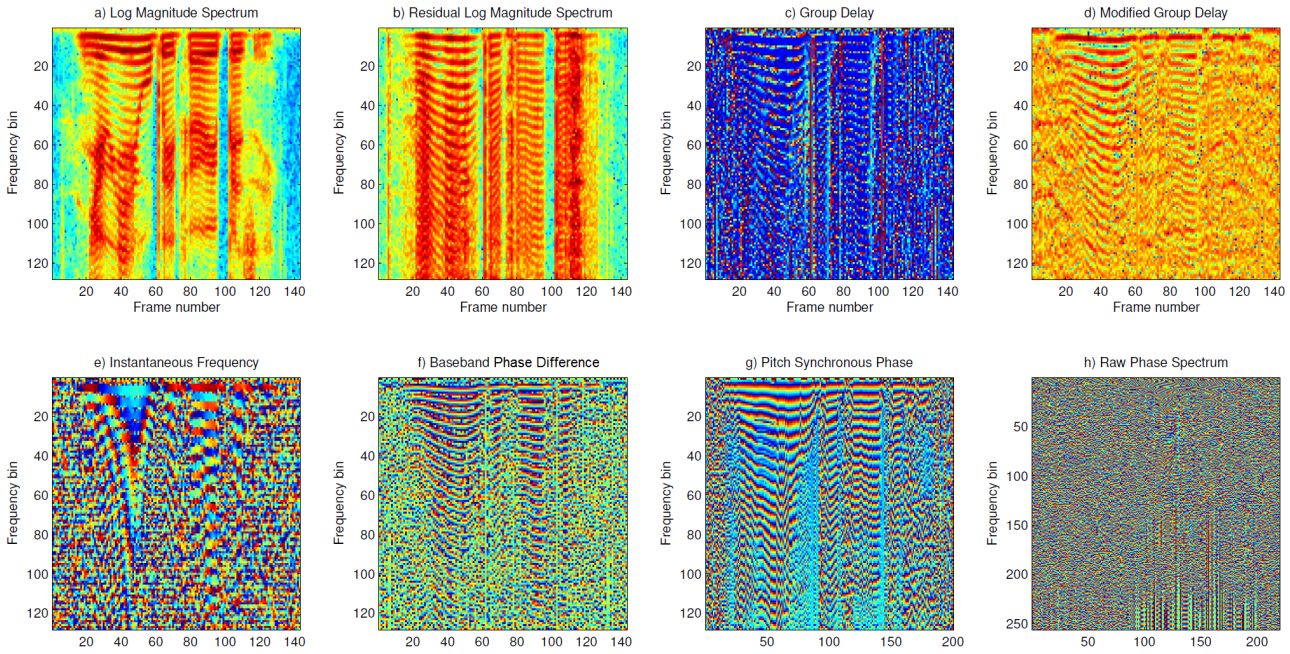

Figure 2: Demonstration of the 7 types of features for utterance D15_1000931, which is a natural speech from the development set. For each feature type, only the low half of the frequency bins are shown.

signals, the formant information are mostly carried by the LPC coefficients, and the LPC residual waveform mainly contains the details of spectrum such as harmonics. It can be observe from Fig. 2b that the formant are gone in the RLMS features.

4.3. Group Delay (GD)

The phase spectrum does not contain stable patterns for spoofing speech detection due to phase warping (see Fig. 2h). Hence, it is necessary to process the phase spectrum to generate useful features for spoofing detection. The first phase based features used in this study is called group delay (Fig. 2c) which is the derivative of phase spectrum along the frequency axis.

$$\mathrm{GD}(t,\omega) = \mathrm{princ}\{\theta(t,\omega) - \theta(t,\omega-1)\} \tag{2}$$

where $\mathrm{princ}(\cdot)$ represents the principal value operator, mapping the input onto $[-\pi,\pi]$ interval by adding integer numbers of 2π. From Fig. 2c, harmonic structure is revealed in the group delay.

4.4. Modified Group Delay (MGD)

The modified group delay (MGD) [21] is an improved version of GD. The MGD feature is computed as follows:

$$\mathrm{MGD} = \frac{\tau(t,\omega)}{|\tau(t,\omega)|}|\tau(t,\omega)|^{\alpha} \tag{3}$$

$$\tau(t,\omega) = \frac{\mathrm{X}_R(t,\omega)\mathrm{Y}_R(t,\omega) + \mathrm{X}_I(t,\omega)\mathrm{Y}_I(t,\omega)}{|\mathrm{S}(t,\omega)|^{2\gamma}} \tag{4}$$

where $\mathrm{Y}(t,\omega)$ is the complex spectrum computed from signal $nx(n)$ and $\mathrm{S}(t,\omega)$ is a smoothed version of $|\mathrm{X}(t,\omega)|$. The subscripts $_R$ and $_I$ denotes real and imaginary parts of the complex spectrum, respectively. The two tuning parameters γ and α are set to 1.2 and 0.4 respectively. By comparing Fig. 2c and d, the MGD has more stable patterns than the GD.

4.5. Instantaneous Frequency Derivative (IF)

While the group delay is the derivative of phase along the frequency axis, the instantaneous frequency is computed as the derivative of the phase along the time axis [21]:

$$\mathrm{IF}(t,\omega) = \mathrm{princ}(\theta(t,\omega) - \theta(t-1,\omega)) \tag{5}$$

By comparing Fig. 2c and e, the IF and GD contain very different patterns, which could provide complementary information for spoofing speech detection.

4.6. Baseband Phase Difference (BPD)

To obtained more stable time-derivative phase based features, we also extract BPD feature [22] as follows:

$$\mathrm{BPD}(t,\omega) = \mathrm{princ}(\mathrm{IF}(t,\omega) - \Omega_t L) \tag{6}$$

where L is the frame shift in terms of number of samples, and $\Omega_t = 2\pi k/N$ is a frequency dependent constant, and N is the FFT length. Fig. 2f shows that the BPD contains different patterns from IF (Fig. 2e).

4.7. Pitch Synchronous Phase (PSP)

Besides the above described phase processing, another way to reveal the patterns in phase spectrum is to use pitch synchronous analysis window. Speech consists of periodic and aperiodic signal. To extract the periodic information, the signal should be framed by using pitch period instead of using fixed frame length. Glottal closure instant (GCI) [23] is used to determine the location of the beginning and the end of each pitch period. Two consecutive pitch periods are joint to form one frame. The overlap between two consecutive frame is set to one pitch period. As pitch period changes along the signal, the overlap size varies according to second pitch period in the frame. The stable periodic patten of PSP could be observed in Fig. 2g.

Table 2: EER obtained by component systems and fused system on development set. The system numbering (a to g) are the same as the numbering in Fig. 3.

Systems		Natural against individual spoofing types					All types
		1	2	3	4	5	
a	LMS	0.347	0.254	0.054	0.054	1.603	0.543
b	RLMS	**0.000**	0.093	**0.039**	0.039	1.456	0.486
c	GD	0.054	**0.054**	**0.039**	**0.000**	**0.161**	**0.114**
d	MGD	1.148	2.311	0.147	0.147	2.311	1.572
e	IF	0.161	0.401	0.147	0.147	0.948	0.428
f	BPD	2.243	4.955	0.401	0.347	5.155	3.431
g	PSP	1.950	1.456	0.093	0.054	1.402	1.345
a-g	Fusion	0.044	0.000	0.000	0.442	0.144	0.001

Table 3: EER (%) on evaluation set by the fused system.

Known attacks					Unknown attacks					Avg.
S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
0.0	0.0	0.0	0.0	0.01	0.01	0.0	0.0	0.0	26.1	2.62

5. Experimental Results

5.1. Results on Development Set

The EER obtained by component systems and their fusion are listed in Table 2 and the DET curves are plotted in Fig. 3. From the EER and the DET curves, there is no obvious correlation between the performance of the features and how obvious their visual patterns are. For example, MGD is considered as more stable phase features than GD, however, the performance of MGD is much worse than that of GD. This could be due to that the patterns shown in Fig.2 may not be necessarily the most useful information for spoofing speech detection. The fusion of the 7 systems produces close to zero EER and its DET curve cannot be seen in Fig. 3.

To understand how complementary the componenet systems are, we add them to the fusion system one by one. At the begining, the fusion set contains the best single system *c*. Then, we add system *e* to the fusion set as it leads to the best EER. The process continues until all component systems are added to the fusion set. The EER obtained at each stage is shown in Fig. 4. It is observed that at most times adding a component system reduce the EER significantly and the lowest EER is obtained when all the 7 component systems are fused. This shows that the component systems, and the features they used, are highly complementary for the task of spoofing speech detection.

5.2. Results on Evaluation Set

The EER on the evaluation set obtained by the fusion systems are shown in Table 3. The EER on both "known" and "unknown" spoofing types are mostly zero or very small, except for type S10 where the EER is 26.1%. On average, the EER on "known" and "unknown" spoofing types are 0.29% and 5.23%, respectively. The good performance on S6-S9 spoofing types show that the trained systems generalized well to some unseen spoofing types. The poor performance on S10 may be due to the 7 types of features do not contain useful information to detect S10 or the MLP classifiers are not tuned to discover the information contained in the features that is useful to differentiate S10 spoofing speech from natural speech.

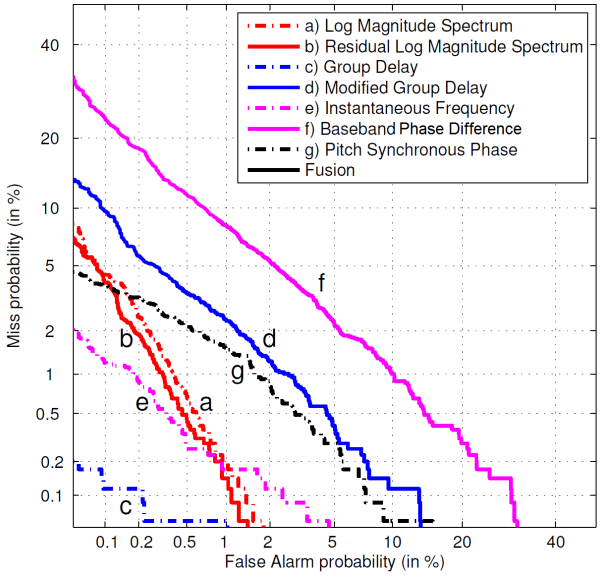

Figure 3: DET curves on development set. The DET curve of the fusion system cannot be seen in the plot as the EER is close to 0.

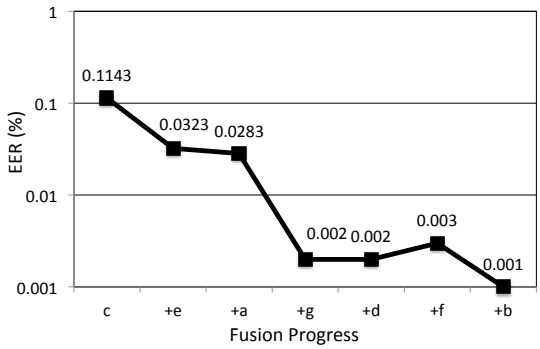

Figure 4: EER at each step of the greedy fusion procedure.

6. Conclusion

In this paper, we described the NTU system for the ASVspoof 2015 spoofing speech detection task. We built 7 component systems using different features. Their scores are averaged to produce the final score for detection. Both high dimensional magnitude and phase based features are used, and long term temporal information up to 0.51s is exploited. To handle the high dimensional features, MLP is used as the classifier and trained to predict posterior probabilities of the incoming sentence being spoofing speech. We observe that EER on most known and unknown spoofing types are zero or very small except for one type of spoofing speech. Future research may be carried out to understand exactly which information are useful for spoofing speech detection and also to improve the generalization capability of the proposed system.

7. Acknowledgements

This work is supported by DSO funded project MAISON DSOCL14045.

8. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[2] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2014.

[3] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[4] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates." in *INTERSPEECH*, 2007, pp. 2053–2056.

[5] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. 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.

[6] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of hmm-based speaker verification systems against imposture using synthetic speech," in *Proc. Eurospeech*, 1999.

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

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

[9] P. L. De Leon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," 2010.

[10] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[11] T. Toda, A. W. Black, and K. 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.

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

[13] Q. Jin, A. R. Toth, A. W. Black, and T. 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.

[14] Q. Jin, A. R. Toth, T. Schultz, and A. 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.

[15] Z. Wu and H. 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] Z. Wu, A. Larcher, K.-A. Lee, E. Chng, T. Kinnunen, and H. Li, "Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints." in *Proc. INTERSPEECH*, 2013, pp. 950–954.

[17] Z. Wu, C. E. Siong, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. INTERSPEECH*, 2012.

[18] Z. Wu, T. Kinnunen, E. S. Chng, H. Li, and E. 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.

[19] Z. Wu, X. Xiao, E. S. Chng, and H. 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.

[20] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," DTIC Document, Tech. Rep., 1997.

[21] L. D. Alsteris and K. 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.

[22] M. Krawczyk and T. Gerkmann, "Stft phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1931–1940, 2014.

[23] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.