SPOOFING DETECTION FROM A FEATURE REPRESENTATION PERSPECTIVE

Xiaohai Tian[1,2]*, Zhizheng Wu*[3]*, Xiong Xiao*[4]*, Eng Siong Chng*[1,2,4] *and Haizhou Li*[1,5]

[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]The Center for Speech Technology Research, University of Edinburgh, United Kingdom
[4]Temasek Laboratories, NTU, Singapore
[5]Human Language Technology Department, Institute for Infocomm Research, Singapore
{xhtian, xiaoxiong, aseschng}@ntu.edu.sg, zhizheng.wu@ed.ac.uk, hli@i2r.a-star.edu.sg

ABSTRACT

Spoofing detection, which discriminates the spoofed speech from the natural speech, has gained much attention recently. Low-dimensional features that are used in speaker recognition/verification are also used in spoofing detection. Unfortunately, they don't capture sufficient information required for spoofing detection. In this work, we investigate the use of high-dimensional features for spoofing detection, that maybe more sensitive to the artifacts in the spoofed speech. Six types of high-dimensional feature are employed. For each kind of feature, four different representations are extracted, i.e. the original high-dimensional feature, corresponding low-dimensional feature, the low- and the high-frequency regions of the original high-dimensional feature. Dynamic features are also calculated to assess the effectiveness of the temporal information to detect the artifacts across frames. A neural network-based classifier is adopted to handle the high-dimensional features. Experimental results on the standard ASVspoof 2015 corpus suggest that high-dimensional features and dynamic features are useful for spoofing attack detection. A fusion of them has been shown to achieve 0.0% the equal error rates for nine of ten attack types.

Index Terms— Spoofing attack, spoofing detection, countermeasure, high-dimensional feature, phase

1. INTRODUCTION

Automatic speaker verification (ASV) system aims to verify the claimed identity based on a given speech signal. For many security systems, the robustness of ASV system against the spoofing attacks is a major concern. However, even the state-of-the-art ASV systems are not designed to sustain spoofing attacks, such as impersonation, replay, speech synthesis and voice conversion, as reviewed in [1]. Specifically, attacks realised by speech synthesis [2] and voice conversion [3], which provide easily accessible ways to generate high quality speech of the target speaker, impose a genuine threat to the ASV systems [4]. To address the threat of these two kinds of synthetic spoofing attacks, one way is to improve the robustness of ASV system by combining the detection process and verification process [5]. An alternative way is to build some standalone detection systems using different countermeasures, and this is also the focus of this paper.

Selecting an effective feature to detect spoofing attacks is important. Considering the phase information is usually neglected by many synthetic techniques, phase-based features are typically used for anti-spoofing, such as modified group delay (MGD) [6, 7, 8, 9], cosine-normalized phase [6, 9], relative phase shift (RPS) [10, 11, 12, 13], cochlear filter cepstral coefficients plus instantaneous frequency (CFCCIF) [14]. Modulation-base features have been used in [15] to detect temporal artifacts. Deep neutral networks have been also used in [16] to extract more discriminant feature. A comparison between different features for synthetic spoofing detection can be found in [17]. However, these are generally low dimensional features, as they have been processed after several stages including feature extraction and dimension reduction. This is also to facilitate the classifiers used, such as Gaussian mixture model (GMM) [18, 7, 15], which is not designed to model the features with high dimension.

As the synthetic techniques typically exploit the low-dimensional features, much detailed information has been abandoned in the spoofed speech. These detailed information may be comprised in the high-dimensional features extracted directly from the Fourier transform. Hence, detailed information carried in high-dimensional or high-frequency features might be useful for spoofing detection. In our previous work [8], seven high-dimensional features, including both magnitude- and phase-based features, are used for synthetic spoofing detection. Even though the system achieves good performance in the synthetic speech detection task, which confirms the effectiveness of the high-dimensional features, it is still not clear how to select a robust feature for spoofing detection.

In this work, we investigate spoofing detection from a feature representation perspective with the following hypotheses:

- The high-dimensional feature is more effective than the low-dimensional feature, as the high-dimensional features contain more detailed information of magnitude and phase;
- The high-frequency region is effective for synthetic spoofing detection, as this region is rarely modelled by speech synthesis or voice conversion;
- The dynamic temporal information is important for spoofing detection, even for high-dimensional features;
- Features derived from different sources are complementary and a fusion of these features can improve detection performance.

To verify these hypotheses, we conduct a series of experiments using six features and the same neural network-based classifier. For each feature, we extract high-dimensional, and corresponding low-dimensional features. We also employ the low- and high-frequency regions of the high-dimensional features to assess their effectiveness.

2. FEATURE REPRESENTATIONS FOR SPOOFING DETECTION

2.1. Feature extraction

Six types of feature are used in this study. We choose these six types because they show good performance [8]. They include two magnitude-based features, namely log magnitude spectrum (LMS) and residual log magnitude spectrum (RLMS); and four phase-based features, namely instantaneous frequency derivative (IF), baseband phase difference (BPD), group delay (GD) and modified group delay (MGD). All the features are evaluated using short-time Fourier transform (STFT) on the speech signal. The speech is sampled at 16KHz and the STFT analysis window is 25ms with 15ms overlap. The Hamming window and direct current (DC) offset is applied on each analysis frame. The magnitude and phase spectrum, extracted directly from the Fourier transform, are further used to generate different magnitude- and phase-based features, respectively. The FFT length is chosen to be 512 and the dimension of all the original features are 256. The features are summarized as follows:

- **LMS**: The log magnitude spectrum feature, which contains the formant information, harmonic structure and all the spectral detail of speech signal. To reduce the dynamic range of the magnitude spectrum, the logarithmic is used. In case of LMS both large and small magnitude frequency components are visible.
- **RLMS**: To reduce the impact of formant information and better analyse the harmonic structure and spectral details, the residual log magnitude spectrum feature is used in this work. The inverse linear predictive coding (LPC) filter is employed to estimate the glottal waveform from the speech signal; then the LMS is extracted from the residual waveform.
- **IF**: Instantaneous frequency [19] is the derivative of the phase along time axis. Therefore, it captures the temporal information of phase. Unlike the original phase spectrum that hardly shows any patterns, there are clear patterns in the IF spectrum, making it possible to be used as a feature.
- **BPD**: Baseband phase difference [20] is a phase feature extracted from baseband STFT, which can also provide a clear pattern to present phase information.
- **GD**: Group delay [21] is a representation of filter phase response, which is defined as the negative derivative of the Fourier transform phase. It is a frame-based feature, used to capture the phase distortion along frequency axis.
- **MGD**: A variation of GD. The modified group delay [21] can obtain a more clear phase pattern than GD. Two factors, α and γ, are used for control the dynamic range of the modified group delay, here the same setting as suggested in [8].

2.2. Feature representation

In order to investigate the effectiveness of the high-dimensional features and relative contribution of the low- and the high-frequency region for synthetic spoofing detection, different feature representations are used in this study, summarized as follows:

- **HD**: The original high-dimensional features which extracted directly from the Fourier transform without dimension reduction, as described in Section 2.1. The dimension of HD feature is 256.
- **LD**: The low-dimensional representation of the HD feature. The Mel-frequency scaled filter banks are computed over the corresponding features to generate 23 filter bank coefficients.
- **HD-LF**: The low-frequency range of the HD feature, consisting of the first half of 128 frequency bins.
- **HD-HF**: The high-frequency range of the HD feature, consisting of the second half of 128 frequency bins.

The effectiveness of temporal information for synthetic spoofing detection has been shown in many works [22, 17]. In [8], we use 51 frames and achieved good performance. Such a high-dimensional feature, however, will increase the difficulty of classifier training. Hence, the dynamic features, including delta and acceleration coefficients, of all the representations noted above are extracted in this study to capture the temporal information, namely HD-D, LD-D, HD-LF-D, HD-HF-D.

2.3. Fusion

Considering the complementarity between different features, a score level fusion is applied. Previous studies [8, 17] shown that single type of feature is usually effective for certain types of artifact, while may not be sensitive to other types of artifact. For example, the main propose the LMS and RLMS features are to capture the artifacts of magnitude, while the IF, BPD, GD and MGD are expected to detect the distortion of phase. Hence, the system fusion is employed to leverage and synergy the merits of different features. To avoid over-fitting to development data, the scores of all systems are simply averaged to produce the final score.

3. EXPERIMENTS

3.1. Experimental setup

The ASVspoof 2015 database [23] is used to assess the performance of different features for synthetic spoofing detection. This database consists of three subsets, including training set, development set and evaluation set. There are totally 10 types of spoofing attack, namely S1 to S10. The training and development sets only contain the spoofing attacks generated by the first five methods (S1-S5); while the evaluation set consists of the spoofing attack generated by both five *known* methods (S1-S5) and five *unknown* methods (S6-S10). Noted that, vocoders and low-dimensional features are used in all the first nine methods (S1-S9)[1]; while waveform-based unit selection is used to generate S10. More details about the database can be found in [23].

Since we focus on feature representations, all the spoofing detectors employed the same neural network (NN) based classifier, as it is capable to model high-dimensional features. The NN consists of one hidden layer with 2,048 sigmoid nodes and is used to perform frame-wise classification. Given testing utterance, the posterior probabilities of all the frames are estimated and averaged over the test utterance.

The equal error rate (EER) is used to evaluate the system performance. The EER is obtained by selecting an operating point which gives the equal miss rate and false alarm rate. In practice, the EERs of each feature for different spoofing attacks are computed using the Bosaris toolkit[2]. For feature-based systems, we only report the results on the evaluation set; while, for each speech synthesis/voice conversion methods of the development set (S1-S5) and the evaluation set (S1-S10), the averaged EERs of the fused systems are reported.

[1] Our system achieved the best performance on S1-S9 in the ASVspoof 2015 challenge.

[2] https://sites.google.com/site/bosaristoolkit/

Table 1. *Averaged EERs (%) of different features on the evaluation set.*

	S1-S5 (Known)						S6-S9 (Unknown)						S10 (Unknown)					
Feature	LMS	RLMS	IF	BPD	GD	MGD	LMS	RLMS	IF	BPD	GD	MGD	LMS	RLMS	IF	BPD	GD	MGD
LD	8.53	5.18	17.79	4.57	19.52	20.70	10.27	7.34	14.28	4.39	19.16	21.57	40.56	47.08	40.58	40.58	**32.8**	45.88
LD-D	3.08	4.72	15.66	3.32	11.16	23.60	3.88	6.45	16.00	2.96	18.87	25.19	38.14	46.52	42.37	**29.59**	39.01	48.31
HD	0.06	0.76	**0.58**	0.72	0.71	0.11	0.06	0.81	**0.74**	0.92	0.48	0.30	41.42	40.79	39.05	37.61	41.83	40.31
HD-D	**0.02**	**0.34**	1.31	**0.10**	**0.05**	**0.00**	**0.01**	**0.36**	1.31	**0.09**	**0.03**	**0.02**	**35.24**	**30.8**	**25.56**	30.67	33.9	**38.54**
HD-LF	5.17	2.10	5.35	5.19	8.49	2.25	3.61	3.55	8.84	8.15	12.56	4.37	42.92	48.76	47.12	47.39	48.61	47.76
HD-LF-D	0.91	1.27	1.22	1.39	1.21	0.29	0.78	2.63	1.59	1.69	2.17	0.45	43.53	48.82	37.77	39.23	46.33	46.02
HD-HF	0.09	1.66	2.84	3.44	4.20	2.24	0.14	2.92	3.03	3.27	4.17	4.00	41.79	41.75	32.63	33.51	38.93	47.59
HD-HF-D	0.03	0.94	1.5	2.59	2.06	0.67	0.05	1.74	1.24	1.90	2.49	4.71	36.68	31.31	27.07	35.51	35.6	43.09

3.2. Results

From our preliminary results, we observe that the results on the development set are similar to that of the S1 to S5 on the evaluation set. Hence, we only report the results on the evaluation set, and the feature-based results are presented in Table 1. Here, we separate the results into three parts, the results of known attacks (S1-S5), the results of unknown attacks using vocoders and low-dimensional features (S6-S9) and the results for unknown attacks generated by waveform (S10).

We first exam the effectiveness of the high-dimensional features. As expected, the high-dimensional (HD) features always outperform its corresponding low-dimensional (LD) representation. Similar phenomenons were observed when dynamic features are included, in particular the HD-D achieves lower error rates than the LD-D. This is due to the low-dimensional feature is not capable in capturing as much detailed information as the high dimensional feature. It also implies that the detailed magnitude and phase information of speech signal contain more artificial cues for synthetic spoofing detection.

We then compare the performance of the low- and the high-frequency regions of the high-dimensional features. The results show that the HD-HF feature always outperforms the corresponding HD-LF feature. This suggests that the high-frequency region of synthetic voice contains more artifacts than the low-frequency counterpart. Low dimensional features, such as Mel-frequency scaled filter banks, have high resolution at low frequency and low resolution at high frequency. They take human perception sensitivity with respect to frequencies into consideration, and therefore are best for speech/speaker recognition. However, they are not designed to be sensitive to the artifacts in synthetic speech signals. Some interesting observations are found by using dynamic features: 1) For magnitude-based features, LMS and RLMS, the high-frequency part (HD-HF-D) outperforms its low-frequency counterpart (HD-LF-D). Especially, for the LMS feature, the HD-HF-D feature (EER of 0.03%) obtain a very close detection accuracy to the HD-D feature (EER of 0.02%), which are much better than the HD-LF-D feature with an EER of 0.91%. This suggests that the high-frequency region is the informative part of the LMS feature. 2) However, for phase-based features, IF, BPD, GD and MGD, while using their dynamic features, the performance of the HD-LF-D and HD-HF-D become close, and the HD-LF-D performs even better in most case.

In general, the dynamic features outperform their static counterparts in most case. This indicates the importance of the temporal information in synthetic spoofing detection. We also note that, comparing to the high-dimensional features, such as HD, HD-HF and HD-LF, less improvement achieve for the low-dimensional (LD) features by including dynamic information. Moreover, comparing to the high-frequency (HD-HF) features, the low-frequency (HD-LF) features obtain more benefits from dynamic features. Overall, the best performance on S1-S5 of evaluation set is obtained by the MGD feature of its high-dimensional with dynamic feature representation (HD-D) achieving an EER of 0.0%.

With that, we now compare the performance across different types of attack. Even though S6-S9 are using similar conversion or synthesis techniques, slightly higher error rates are observed for S6-S9 than that of S1-S5. This is because S1-S5 attacks are available for training, while S6-S9 are only available during evaluation. However, due to the vocoders and low-dimensional features used for generation are similar, the difference between them is not as large as expected. Again, the MGD feature with its HD-D representation achieves good performance on this set (EER of 0.02%). This implies the effectiveness of phase information. The LMS feature with the HD-D representation performs best with an EER of 0.01%.

For the unknown attacks (S10), the performance degrades dramatically. This is due to the fact that all the features used in our systems aim for detecting the artifacts of either magnitude or phase. However, S10 attacks are generated by unit selection algorithm, which use original waveforms directly. Hence, the artifacts only appear in the transition positions, which increase the difficult for detection.

3.3. Fusion results

The results of fused systems on both development and evaluation sets are presented in Table 2. From the results, we have following observations. First, comparing to the systems using single feature, the EERs of the fused system are reduced significantly on both development and evaluation sets. Particularly, for the fused system of HD-D, all the spoofing attacks in development set and most of spoofing attacks in evaluation set (S1-S9) are detected with EERs of 0.0%. Second, all the fused systems fail to detect S10 generating by waveform concatenation. We know that in unit selection most of the speech samples are actually natural speech. The artifacts only exist between the units. Third, again, the results confirm the effectiveness of high-dimensional features, high-frequency information and dynamic features for spoofing attacks detection. Finally, we confirm that all the results are consistent with [8]. Specifically, for the development set and (S1-S9) of evaluation set, our system outperforms the previous system. The best results is achieved by the fusion system of HD-D with the EER of 2.78%

4. DISCUSSIONS

4.1. Magnitude vs phase features

From the perspective of voice conversion and speech synthesis, artifacts are introduced in both magnitude and phase domains. For mag-

Table 2. *EERs (%) of fused system on both development and evaluation sets. Pre-system indicates the results of our previous work [8].*

	Development set						Evaluation set										
Feature	S1	S2	S3	S4	S5	Average	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Average
LD	0.87	2.78	1.06	0.84	1.20	2.55	0.85	2.22	0.61	0.61	7.74	9.38	2.41	0.41	0.81	41.38	6.64
LD-D	0.45	1.83	0.84	0.67	3.98	1.56	0.42	2.07	0.51	0.53	4.98	6.87	1.51	0.27	0.34	40.16	5.77
HD	**0.00**	**0.00**	**0.00**	**0.00**	0.02	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.03	0.02	**0.00**	**0.00**	**0.00**	37.42	3.75
HD-D	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**27.79**	**2.78**
HD-LF	2.41	1.83	0.03	0.07	1.73	1.21	1.34	1.44	0.04	0.08	1.54	1.77	1.45	0.26	1.05	45.95	5.49
HD-LF-D	0.25	0.23	0.01	0.02	0.24	0.15	0.10	0.10	0.00	0.01	0.17	0.22	0.18	0.01	0.07	41.90	4.28
HD-HF	0.04	**0.00**	**0.00**	0.01	0.30	0.07	0.03	0.04	0.02	0.03	0.27	0.29	0.01	**0.00**	0.01	36.20	3.69
HD-HF-D	**0.00**	**0.00**	**0.00**	**0.00**	0.12	0.02	**0.00**	**0.00**	**0.00**	**0.00**	0.13	0.07	**0.00**	**0.00**	**0.00**	28.82	2.90
Pre-system	0.04	0.00	0.00	0.44	0.14	0.12	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	26.10	2.62

nitude, the artifacts are mainly due to over-smoothing in both temporal and frequency domains. Over-smoothing in temporal structure lowers the variation of the spectral trajectory for synthetic speech. While over-smoothing in frequency domain can result in missing of spectral details. These phenomenons have been intensively reported in both voice conversion [3] and speech synthesis [2].

Similarly, for phase, even though it has been shown the usefulness for speech perception [24], it is very hard to find stable patterns for modelling. Hence, in most vocoders, the original/natural phase is discarded and replaced with a minimum phase, thus, leaving obvious artifacts.

The success of magnitude and phase-based features in spoofing detection confirms that the artifacts exist in both these two parts.

4.2. High-dimensional vs low-dimensional features

For magnitude, low-dimensional features are commonly used in voice conversion and speech synthesis. However, the limitation of low-dimensional features for capturing the detailed spectral information is reported in [25, 26]. It was noticed that the use of low-dimensional features result in loss of spectral details in the synthetic speech signals. Hence, there should be more artifact cues in spectral details for synthetic spoofing detection. This is confirmed by the results shown in previous section, which indicate that the low-dimensional features perform much worse than the high-dimensional counterparts.

For phase-based features, such as IF, GD and MGD, the performances of their low-dimensional (LD) representations are much worse than their HD counterparts. Although, after processing, some patterns are observed in the HD phase features, in general, the patterns are still not as clear as those of magnitude-based features. The patterns become blurred after dimension reduction using filter banks.

4.3. High-frequency vs low-frequency features

In general, low-dimensional features, used in most speech synthesis and voice conversion approaches, model the low-frequency of speech information more accurately than its high-frequency counterpart. Additionally, due to the smoothing effect of modelling, the spectral details can be further lost. Hence, with magnitude-based features, artifacts are easier to find in the high-frequency region. Our results, in Section 3, confirm the effectiveness of the high-frequency features. Similar conclusion can also be found in [17].

The situation is a bit different in phase-based features. As not modelled in speech synthesis or in voice conversion, the phase features do not contain the smoothing effects of modelling and the issues concerning low-dimensional features. However, the phase generated by vocoder is still different from that of natural speech. Different to the results of magnitude-based features, with the phase-based features implemented with dynamic features, the low-frequency (HD-LF) outperforms the high-frequency (HD-HF). But, the performance of high-dimensional features (HD) is much better than either the low-frequency or the high-frequency. This suggests that both the low- and the high-frequency of phase-based features contribute to the synthetic spoofing detection.

4.4. Difficulty in detecting unit selection-based attacks

Similar to all the other systems submitted to the ASVspoof 2015 challenge [23], our system doesn't perform as expected in S10, which is an unit-selection based attack. The unit-selection based attack is produced by concatenating the time-domain waveform directly without any vocoding and feature extraction techniques, which doesn't carry much artifacts from the perspective of feature representations.

According to the human perception study [27], human listeners can easily detect the unit-selection attack, as human ear can detect the artifacts of the discontinuity at the waveform concatenating points. We will investigate techniques to detect such discontinuities for spoofing detecting in the future work.

5. CONCLUSIONS

In this paper, we investigate the high-dimensional features and the high-frequency information to discriminate synthetic spoofing attacks and natural speech. The experiments results show that,

- the high-dimensional features outperform its low-dimensional representations;
- the high-frequency and the low-frequency information are complementary in discriminating synthetic spoofing attacks and natural speech;
- the dynamic features are useful for synthetic spoofing attacks detection;
- fusion of different systems improve the detection performance.

Additionally, by applying the score level system fusion, all the spoofing attacks using vocoders and low-dimensional features are successfully detected. While poor performance is observed for unit selection-based spoofing attacks. Therefore, in future, we will investigate more robust feature to detect such kind of spoofing attacks.

Acknowledgment This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme Office. This work is also supported by the DSO funded project MAISON DSOCL14045, Singapore.

6. REFERENCES

[1] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[2] Heiga Zen, Keiichi Tokuda, and Alan W Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[3] Yannis Stylianou, "Voice transformation: a survey," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[4] Z Wu, A Khodabakhsh, C Demiroglu, J Yamagishi, D Saito, T Toda, and S King, "SAS: A speaker verification spoofing database containing diverse attacks," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[5] Elie Khoury, Tomi Kinnunen, Aleksandr Sizov, Zhizheng Wu, and Sébastien Marcel, "Introducing i-vectors for joint anti-spoofing and speaker verification," in *Proc. INTERSPEECH*, 2014.

[6] 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.

[7] 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.

[8] Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Eng Siong Chng, and Haizhou Li, "Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge," in *Proc. INTERSPEECH*, 2015.

[9] Yi Liu, Yao Tian, Liang He, Jia Liu, and Michael T Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," in *Proc. INTERSPEECH*, 2015.

[10] Phillip L De Leon, Inma Hernaez, Ibon Saratxaga, Michael Pucher, and Junichi Yamagishi, "Detection of synthetic speech for the problem of imposture," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4844–4847.

[11] Jon Sanchez, Ibon Saratxaga, Inma Hernaez, Eva Navas, and Daniel Erro, "The AHOLAB RPS SSD spoofing challenge 2015 submission," in *Proc. INTERSPEECH*, 2015.

[12] Longbiao Wang, Yohei Yoshida, Yuta Kawakami, and Seiichi Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proc. INTERSPEECH*, 2015.

[13] Jon Sanchez, Ibon Saratxaga, Inma Hernaez, Eva Navas, Daniel Erro, and Tuomo Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 810–820, 2015.

[14] Tanvina B Patel and Hemant A Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Proc. INTERSPEECH*, 2015.

[15] 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.

[16] Nanxin Chen, Yanmin Qian, Heinrich Dinkel, Bo Chen, and Kai Yu, "Robust deep feature for spoofing detectionthe SJTU system for ASVspoof 2015 challenge," in *Proc. INTERSPEECH*, 2015.

[17] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. INTERSPEECH*, 2015.

[18] 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.

[19] 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.

[20] Michal Krawczyk and Timo Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, 2014.

[21] Bayya Yegnanarayana and Hema A Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Transactions on Signal Processing*, vol. 40, no. 9, pp. 2281–2289, 1992.

[22] Xiaohai Tian, Steven Du, Xiong Xiao, Haihua Xu, Eng Siong Chng, and Haizhou Li, "Detecting synthetic speech using long term magnitude and phase information," in *Proc. IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*. IEEE, 2015, pp. 611–615.

[23] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. INTERSPEECH*, 2015.

[24] Kuldip K Paliwal and Leigh D Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153–170, 2005.

[25] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[26] Cassia Valentini-Botinhao, Zhizheng Wu, and Simon King, "Towards minimum perceptual error training for DNN-based speech synthesis," in *Proc. INTERSPEECH*, 2015.

[27] Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi, "Human vs machine spoofing detection on wideband and narrowband data," in *Proc. INTERSPEECH*, 2015.