Combining Phase-based Features for Replay Spoof Detection System

Kantheti Srinivas[1], *Rohan Kumar Das*[2] *and Hemant A. Patil*[1]

[1]Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT),
Gandhinagar, India
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
{srinivas_kantheti,hemant_patil}@daiict.ac.in, rohankd@nus.edu.sg

Abstract

Automatic Speaker Verification (ASV) systems are developed to verify the claimed identity of a speaker based on speech samples. The technological advances have given pathways to practical ASV systems that showcase the threat towards spoofing attacks. Replay is one of the spoofing attacks where the ASV systems are fooled with pre-recorded speech samples of a target speaker. In this context, both magnitude-based and phase-based spectral features get affected by the quality of intermediate devices and their environments. There have been only a few studies reported to detect the replay attacks based on the phase features. In this paper, we explore the relative significance of various phase-based features for detecting replay attacks. The magnitude-based features are chosen to perform score-level fusion with phase-based features to capture the possible complementary information. Among various possible combinations of magnitude and phase-based features, we obtain 12.25 % as the best Equal Error Rate (EER) which is less than that obtained with individual feature set, while the score-level fusion of phase-based features gave an EER of 13.14 % on the evaluation set of ASVspoof 2017 version 1 database.

Index Terms: ASV, spoofing, relative phase shift, modified group delay, linear prediction residual phase.

1. Introduction

The recent technological advances in speech processing lead to the use of an Automatic Speaker Verification (ASV) system for voice biometric applications. There are two types of ASV systems, namely, text-independent and text-dependent. The ASV system performance should be robust against variabilities occurred from microphone, transmission channel, acoustic noise, etc. However, the ASV systems are vulnerable to various spoofing attacks, namely, impersonation, identical twins, replay, speech synthesis and voice conversion. The replay attacks use pre-recorded speech samples of a target speaker to get access of the system [1]. These kind of attacks do not require sophisticated signal processing techniques which is required for the synthetic speech attacks. The replay spoofing is of two types, namely, closer and far-field recording [2]. The second edition of ASVspoof Challenge was organized focusing on replay spoof attacks in INTERSPEECH 2017 [3]. The vulnerability of replay attacks increases due to use of high quality intermediate devices and clean acoustical conditions, which exhibits similar characteristics of natural speech signal of genuine speaker [3].

Various countermeasures were proposed at the front-end and back-end for replay Spoof Speech Detection (SSD) task. The spectral bitmaps were used for replay spoof detection in a text-dependent speaker verification system [4]. The average spectral bitmaps and cosine kernel score techniques were proposed for replay detection in [2]. The detection of replay spoofing was attempted by analyzing the acoustical characteristics of a given utterance and copy detection algorithm were used to predict the recordings of a genuine utterance [3]. The spectral ratio and modulation index were proposed for the detection of far-field recorded signal [5]. The authors of [6] detected the replay attacks based on the acoustical characteristics.

The importance of spectral statistics features and their analysis on the effect of channels by deep learning approaches were studied in [7]. The ReliefF algorithm was used to reduce overfitting and variance in features with Support Vector Machine (SVM) classifier [8]. The deep learning approaches, such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and stacking of both at front-end and i-vector-based SVM classifier at backend were also studied in [9]. Various deep learning approaches, such as CNN, RNN and Long Short-Term Memory were also used for detecting replay attacks [10, 11, 12]. The analysis of various magnitude spectrum-based features and their score-level fusion was studied in [13]. The high frequency regions were found to be more significant than the low frequency regions [14]. The source-based features, such as epoch features, Peak-to-Side-lobe Ratio of Mean and Skewness and cepstral features were studied in [15]. The Variable length Teager Energy operator-Energy Separation Algorithm-Instantaneous Frequency Cosine Coefficients (VESA-IFCC) were proposed to discriminate the genuine and replay speech [16].

For the past few decades, phase-based features are not widely explored in speech applications, however, the phase also contains useful information of speech signal. The overview of various phase-based features is presented in [17]. The phase-based features are investigated less for replay detection [18]. In this work, we have explored various phase-based features, such as Relative Phase Shift (RPS) feature [19], Modified Group Delay Cepstral Coefficients (MGDCC) [20], and Linear Prediction Residual Phase Cepstral Coefficients (LPRPCC) [21] for replay spoof detection. To extract the possible complementary information of these phase-based features, we have fused them at the score-level. In addition, we compared our results with various magnitude spectrum-based features, such as Constant-Q Cepstral Coefficients (CQCC) [22], Mel Frequency Cepstral Coefficients (MFCC) [23] and Linear Frequency Cepstral Coefficients (LFCC) [24].

2. Feature Extraction

2.1. Discrete Cosine Transform-Linear-Relative Phase Shift (DCT-Linear-RPS)

Relative Phase Shift (RPS) feature is computed from the instantaneous phase of the speech signal, which is the representation of initial phase shift differences with reference to the fundamental frequency (F_0). The phase information represents two properties of the signal, namely, waveform shape, and time synchronicity [25]. The difference of harmonic phase components (i.e., RPS) decides the waveform shape and contains the harmonic phase information of the speech signal. The harmonic analysis decomposes the speech signal into periodic and aperiodic components [25]. The periodic components are represented as the sum of harmonic sinusoidal components with an integer multiple of fundamental frequency (F_0) in each frame and it is represented as [25]:

$$s(t) = \sum_{i=1}^{K} a_i \sin(\phi_i(t)), \tag{1}$$

where $\phi_i(t) = 2\pi i F_0 t + \theta_i$, i is the harmonic index, a_i and θ_i are the amplitude and initial phase of i^{th} harmonic component, respectively. The RPS is phase shift of the instantaneous phase of every harmonic and fundamental instantaneous phase of the harmonic component at a particular time instant t_c of the period of the signal. At this time instant t_c, the phase difference (RPS) is constant while the waveform shape is constant (i.e., $\phi_1(t_c) = 0$) [26]. RPS is shown in Eq. (2) [19, 26]:

$$\psi_i(t) = \phi_i(t_c) = \phi_i(t) - i\phi_1(t), \tag{2}$$

where $\phi_i(t)$ is the instantaneous phase of i^{th} harmonic component, and $\phi_1(t)$ is fundamental harmonic component at any time instant t. The advantage of RPS is the removal of linear phase term, which is the cause of constant phase wrapping in instantaneous phase. Hence, the RPS feature emerges smoothly in time-axis [27]. If RPS is zero then it is known as cosine phase features, otherwise, the RPS feature contains random phase values $[-\pi, \pi]$ [26].

The number of RPS values depends upon a number of harmonic values which has higher dimension and wrapping discontinuities. This motivated us to apply parameterization [19, 28] to RPS feature in each frame of a speech signal. The first step in parameterization is unwrapping process which is an ambiguous operation that generates the data different from the RPS values, so the difference of unwrapped RPS values should be taken. The difference of unwrapped RPS values are filtered with linear triangular filterbank and in addition, the average of differentiated unwrapped RPS values is appended. Furthermore, Discrete Cosine Transform (DCT) is taken to obtain DCT-Linear-RPS feature vector.

2.2. Linear-Frequency Modified Group Delay Cepstral Coefficients (LFMGDCC)

To extract LFMGDCC feature set, the speech signal $s(n)$ is analyzed with the help of Fourier Transform (FT) of short duration segments. The FT of $s(n)$ is represented in magnitude and phase form as:

$$S(\omega) = |S(\omega)|e^{j\phi(\omega)}. \tag{3}$$

To extract the phase information $\phi(\omega)$ from the speech signal, the Group Delay (GD) function is used. It is defined as the negative derivative of FT phase (i.e., $\phi(\omega)$) w.r.t. frequency 'ω' as shown in Eq. (4) [29]:

$$\tau(\omega) = -\frac{d}{d\omega}\phi(\omega), \quad \tau(\omega) = -\text{Im}[\frac{d}{d\omega}\log(S(\omega))]. \tag{4}$$

The Eq. (4) requires unwrapping in phase before differentiation. To avoid this complex task, we compute the GD function using the property of FT as shown in Eq. (5) [29]:

$$\tau(\omega) = \frac{S_r(\omega)Q_r(\omega) + S_i(\omega)Q_i(\omega)}{|S(\omega)|^2}, \tag{5}$$

where $Q(\omega)$ represents the FT of $ns(n)$, i - imaginary part and r - real part. The spikes and pitch periodicity effects are introduced in GD function [29]. The cause of producing spikes here is that the denominator term tend to be very small if zeros of $S(\omega)$ lie close to the unit circle in the z-plane [29]. This problem cannot be eliminated by normal smoothing techniques. To avoid these effects, the GD function is modified, which is known as Modified Group Delay (MGD) function. The spikes can be suppressed by cepstral smoothing of denominator term $|S(\omega)|$ and with the help of tuning parameters (ρ, γ). The parameters ρ and γ are tuned to which depends on the problem of study at hand [29, 30]. Using this parameters, we can do normalization where we can reduce the problem of spikes. The MGD function is shown in Eq. (6):

$$\tau(\omega) = \frac{S_r(\omega)Q_r(\omega) + S_i(\omega)Q_i(\omega)}{|S_c(\omega)|^{2\rho}}, \; \tau_m(\omega) = \frac{\tau(\omega)}{|\tau(\omega)|}|\tau(\omega)|^{\gamma}, \tag{6}$$

where $|S_c(\omega)|$ is smooth cepstral value of $|S(\omega)|$. The tuning parameters ρ, γ varies between 0 to 1. The details of Algorithm for Mel Frequency MGDCC (MFMGDCC) extraction is given in [31]. In addition, the Linear Frequency MGDCC (LFMGDCC) is extracted here with the help of a linear frequency scale.

2.3. Linear Prediction Residual Phase Cepstral Coefficients (LPRPCC)

The Linear Prediction (LP) residual phase of a speech signal represents the excitation source components [21]. It is computed with the help of analytic signal obtained from the LP residual. The LP analysis predicts the current speech sample $s(n)$ from the past 'N' samples. The estimated sample $\hat{s}(n)$ is represented as:

$$\hat{s}(n) = -\sum_{k=1}^{N} b_k s(n-k), \tag{7}$$

where N is the prediction order, and b_k is k^{th} LP coefficient (LPC) from LP analysis. The LP residual $r(n)$ is the difference of actual sample and estimated sample as shown in Eq. (8) [32]:

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{N} b_k s(n-k). \tag{8}$$

The $r(n)$ values are large around Glottal Closure Instants (GCIs) in voiced segments of speech signal [32]. The occurrence of fluctuations in amplitudes causes no proper information to extract from the short-term segments of LP residual. Hence, the LP residual phase features were proposed [32, 33]. The LPRPCC features are extracted from the analytic signal of $r(n)$ as shown in Eq. (9):

$$r_a(n) = r(n) + jr_h(n), \tag{9}$$

where $r_a(n)$ is the analytic signal of LP residual and $r_h(n)$ is the Hilbert transform of $r(n)$. The $r(n)$ phase is the cosine of the phase of an analytic signal $r_a(n)$ as shown in Eq. (10):

$$\cos(\theta(n)) = \frac{r(n)}{\sqrt{r(n)^2 + r_h(n)^2}}. \tag{10}$$

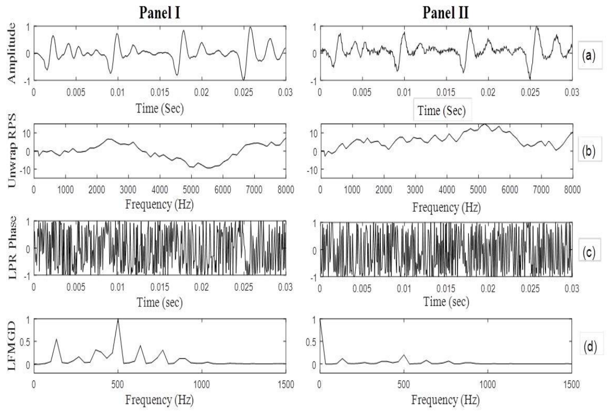

Figure 1: *Phase-based features of (a) voiced speech segment, (b) unwrapped RPS features, (c) LP residual phase, (d) linear frequency scale MGD. Panel I and Panel II are natural and replay signal, respectively.*

Table 1: *Details of magnitude and phase-based features*

Feature set	SSD system
CQCC	Magnitude feature1 (C1)
MFCC	Magnitude feature2 (C2)
LFCC	Magnitude feature3 (C3)
DCT-Linear-RPS	Phase feature1 (P1)
LFMGDCC	Phase feature2 (P2)
LPRPCC	Phase feature3 (P3)

Apply DCT to cosine phase features computed via Eq. (10) for energy compaction, and then Δ, $\Delta\Delta$ coefficients are appended to form LPRPCC feature vector.

Figure 1 shows the behavior of the phase-based features for genuine and replay voiced speech segment. It can be observed that there is some difference in phase-based features of genuine and replay speech segments, even though both the segments are similar. In addition, we find that the phase-based features are sensitive to the noise due to replay. Hence, may be helpful to use phase-based features for replay SSD task, because most of the part in replay speech contains noise due to replay mechanism.

3. Experimental Results

3.1. Database and Classifier

The experiments were performed on ASVspoof 2017 version 1 database. The sampling rate is 16 kHz with 16-bits resolution per sample. The database consists of three subsets, namely, Train, Development and Evaluation. The train set is given for learning the Gaussian mixture model (GMM) corresponding to genuine and replay speech [3]. The development set is used for learning the parameters for fusion and then they are applied on the evaluation set. The details of the database is given in [3].

The baseline system with CQCC features has 96 number of bins per octave and the minimum and maximum frequency of 15 Hz and 8 kHz (half of the sampling frequency due to Shannon's sampling theorem). The first 30 coefficients are retained using DCT and Δ, $\Delta\Delta$ coefficients are appended to form *90*-dimensional (D) feature vector. The MFCC feature set is extracted with the help of Hamming window of 20 ms duration with 10 ms shift and 40 subband filters in Mel triangular filterbank results into *39*-D feature vector. The LFCC feature is extracted with Hamming window of 25 ms and 10 ms shift with the 60 linear triangular subband filters which results into a *180*-D feature vector.

Table 2: *Results (in EER %) of Individual magnitude and phase-based features*

SSD system	EER (%)	
	Dev	Eval
C1 (Baseline)	11.89	28.92
C2	11.21	31.30
C3	10.28	16.80
P1	29.50	23.50
P2	19.94	21.66
P3	34.08	23.17

Table 1 shows the various features used in replay SSD systems. In phase-based features, the window duration and shift plays a crucial role and also the shape of window function as reported in [34]. The P1 feature set is extracted from the preprocessed (pre-emphasis) speech signal. The pre-processing is applied for boosting the high frequency components of speech signal. The rectangular window is used for segmentation and 48 subband filters are used in linear triangular filterbank. The first 13 static coefficients are appended with Δ, $\Delta\Delta$ coefficients to form *39*-D feature vector. The Cepstral Mean Normalization (CMN) is applied for channel compensation and the 1024 Gaussian components are used in GMM.

The P2 feature set is extracted similarly using Hanning window of duration 25 ms, 10 ms shift and the tuning parameters ρ and γ are tuned to 0.4 and 0.1, respectively. Total 40 number of linear triangular filters in filterbank is used, and first 13 coefficients are retained to form static, Δ and $\Delta\Delta$ are appended to form *39*-D feature vector. The CMN is also applied to normalize the features. The P3 feature set is then extracted considering Hanning window is used for segmentation of duration 25 ms and 15 ms shift with 1024 number of GMM components. The first 13 static coefficients are retained and appended with Δ, $\Delta\Delta$ to obtain 39-D feature vector, where CMN is applied.

Table 2 shows the results for the three magnitude features and the three phase-based features discussed in this work. It is observed that the phase-based features perform better to the C1 and C2-based magnitude features on comparing the evaluation set. However, it is to be mentioned that their performance is on development set do not follow the same trend. We believe that the fusion of magnitude and phase features would complement each other that will contribute towards improving the overall performance for spoof detection.

The results of score-level fusion of different feature combinations are reported in Table 3. The score-level combination of the features was found to be positive which improved the individual feature performance. For the magnitude feature C1, the phase feature P3 gave the best performance on fusion for both development as well as evaluation set. On the other hand, for the magnitude feature C2, the phase feature P2 was found to be most effective. Similarly, the P1 phase feature resulted as the most suitable combination for magnitude feature C3 on the evaluation set. Comparing the results of different combination pairs of magnitude and phase features, the combination of C3 and P1 provides the best performance.

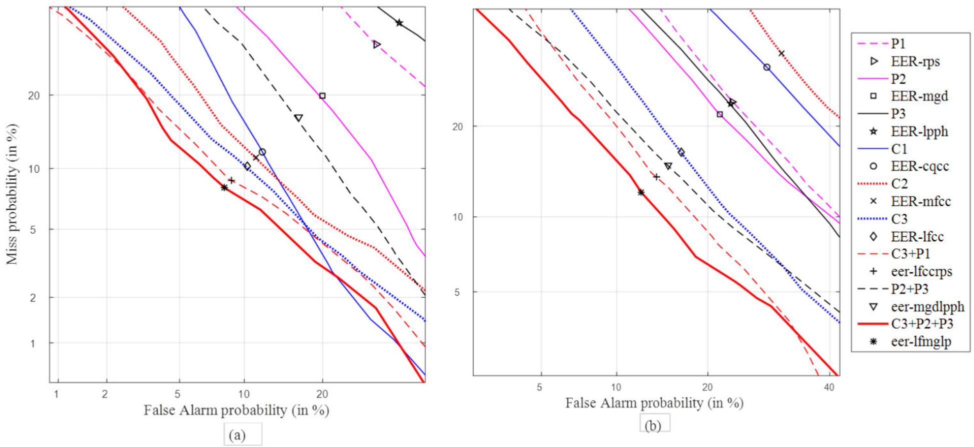

Figure 2: *DET curves of various replay SSD systems on (a) development dataset, and (b) evaluation dataset.*

Table 3: *Results of score-level fusion (in % EER) on development and evaluation set*

SSD System	Dev	Eval
C1 + P1	11.24	18.56
C1 + P2	9.55	19.91
C1 + P3	9.43	18.35
C2 + P1	10.25	21.07
C2 + P2	9.0	20.73
C2 + P3	10.83	22.03
C3 + P1	**8.79**	**13.81**
C3 + P2	9.30	14.22
C3 + P3	8.57	14.31
P1 + P2	14.28	16.48
P1 + P3	25.40	19.56
P2 + P3	**16.47**	**15.10**

+ : indicates score-level fusion

We then investigate the complementary information captured by one type of phase features to the other by having score-level fusion. The performance for their combination is reported in Table 3. It is observed that the we obtain a gain with their score-level fusion. The phase features P1 and P3 contain cosine phase information and hence, the complementariness is less compared to the other phase feature P2. This results in comparatively better performance when P1 or P3 is fused with P2 phase features. In addition, the combination of P2 and P3 phase features provides the best result among them on the evaluation set.

Table 4 shows the results of score-level fusion of three feature combinations for the detection of replay attacks. The fusion of three phase-based features (P1+P2+P3) gives 12.32 % on development dataset and 13.14 % on evaluation dataset. However, the best combination is obtained when both magnitude and phase features are considered. The combination (P2+P3+C3) yields the best results as can be observed from Table 4. The Detection Error Trade-off (DET) [35] curves in Figure 2 indicate that the performance of phase-based features fusion (P2+P3) compared with that of fusion between magnitude-based and phase-based features (C3+P1). Similarly, the performance of the fusion of three feature sets between magnitude and two phase-based feature sets (P2+P3+C3) is better compared with the performance of (P1+P2+P3) feature set, individual phase-based and magnitude-based features in both development and evaluation set.

Table 4: *Results of score-level fusion of three feature sets (in % EER) on development and evaluation set*

SSD System	Dev	Eval
P3 + P2 + P1	**12.32**	**13.14**
P3 + P2 + C3	**8.17**	**12.25**
P3 + P2 + C1	7.78	13.97
P3 + P2 + C2	8.40	14.99
P3 + P1 + C3	7.97	12.63
P3 + P1 + C1	8.75	15.84
P3 + P1 + C2	9.49	17.77
P2 + P1 + C3	8.09	12.44
P2 + P1 + C1	8.77	14.51
P2 + P1 + C2	7.54	16.18

4. Summary and Conclusions

This study investigated the relative significance of various phase-based features for replay SSD task. It is observed that the performance of linear frequency scale-based MGD feature set is better compared to other phase-based features for replay SSD task. Furthermore, these features are fused at score-level with each other and with existing spectral (segmental) features, such as CQCC, MFCC and LFCC. Fusion with only phase-based features without using magnitude features performed better in evaluation dataset that is close to the performance obtained with magnitude features and poorer than the performance of magnitude spectral features on the development dataset. In addition, the fusion of LFCC with DCT-Linear-RPS performed better in both development and evaluation sets. The studies showed that phase-based features contain complementary information from the magnitude features. Hence, their fusion resulted to improve the performance. The future work will focus towards investigating deep learning architectures for phase-based features to detect replay attacks.

5. Acknowledgements

The first and the last authors would like to thank the authorities of DA-IICT, Gandhinagar and MeitY, Govt. of India for their kind support to carry out this work. The work of second author is supported by Programmatic Grant No. A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain).

6. References

[1] 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.

[2] A. Paul, R. K. Das, R. Sinha, and S. R. M. Prasanna, "Countermeasure to handle replay attacks in practical speaker verification systems," in *IEEE International Conference in Signal Processing and Communications (SPCOM), Bangalore, India*, 2016, pp. 1–5.

[3] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASV Spoof 2017 Challenge: Assessing the limits of replay spoofing attack detection," in *INTERSPEECH, Stockholm, Sweden*, 2017, pp. 2–6.

[4] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *IEEE Annual Summit and Conference of Asia-Pacific Signal and Information Processing Association (APSIPA)*, Angkor Wat, Cambodia, 2014, pp. 1–5.

[5] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *IEEE International Carnahan Conference on Security Technology (ICCST), Barcelona, Spain*, 2011, pp. 1–8.

[6] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *IEEE International Conference of Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, 2014, pp. 1–6.

[7] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "Presentation attack detection using long-term spectral statistics for trustworthy speaker verification," in *IEEE International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, 2016, pp. 1–6.

[8] X. Wang, Y. Xiao, and X. Zhu, "Feature selection based on CQCCs for automatic speaker verification spoofing," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 32–36.

[9] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 82–86.

[10] K. R. Alluri, S. Achanta, S. R. Kadiri, S. V. Gangashetty, and A. K. Vuppala, "SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge," in *INTERSPEECH, Stockholm, Sweden*, 2017, pp. 107–111.

[11] W. Cai, D. Cai, W. Liu, G. Li, and M. Li, "Countermeasures for automatic speaker verification replay spoofing attack: On data augmentation, feature representation, classification and fusion," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 17–21.

[12] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using DNN for channel discrimination," in *INTERSPEECH, Stockholm, Sweden*, 2017, pp. 97–101.

[13] R. Font, J. M. Espın, and M. J. Cano, "Experimental analysis of features for replay attack detection–results on the ASV Spoof 2017 Challenge," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 7–11.

[14] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio replay attack detection using high-frequency features," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 27–31.

[15] S. Jelil, R. K. Das, S. R. M. Prasanna, and R. Sinha, "Spoof detection using source, instantaneous frequency and cepstral features," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 22–26.

[16] H. A. Patil, M. R. Kamble, T. B. Patel, and M. Soni, "Novel variable length Teager energy separation based instantaneous frequency features for replay detection," in *INTERSPEECH, Stockholm, Sweden*, 2017, pp. 12–16.

[17] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.

[18] R. K. Das and H. Li, "Instantaneous phase and excitation source features for detection of replay attacks," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Honolulu, Hawaii, 2018.

[19] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. 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.

[20] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 190–202, 2007.

[21] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 309–319, 1979.

[22] M. Todisco, H. Delgado, and N. W. Evans, "Articulation rate filtering of CQCC features for automatic speaker verification." in *INTERSPEECH*, San Franscisco, USA, 2016, pp. 3628–3632.

[23] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing Mel-frequency cepstral coefficients on the power spectrum," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Salt Lake City, Utah, 2001, pp. 73–76.

[24] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus Mel frequency cepstral coefficients for speaker recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, New York, USA, 2011, pp. 559–564.

[25] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez, "Simple representation of signal phase for harmonic speech models," *Electronics Letters*, vol. 45, no. 7, pp. 381–383, 2009.

[26] I. Saratxaga, I. Hernaez, M. Pucher, E. Navas, and I. Sainz, "Perceptual importance of the phase related information in speech," in *INTERSPEECH*, Portland, USA, 2012, pp. 1448–1451.

[27] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 1, pp. 1–15, 2014.

[28] I. Hernáez, I. Saratxaga, J. Sanchez, E. Navas, and I. Luengo, "Use of the harmonic phase in speaker recognition," in *INTERSPEECH*, Florence, Italy, 2011, pp. 2757–2760.

[29] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Hong Kong, China, 2003, pp. 68–71.

[30] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, British Colombia, Canada, 2013, pp. 7234–7238.

[31] D. Zhu and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Quebec, Canada, 2004, pp. 125–128.

[32] K. S. R. Murty and B. Yegnanarayana, "Combining evidences from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, 2006.

[33] C. Hanilçi, "Speaker verification anti-spoofing using linear prediction residual phase features," in *European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, 2017, pp. 96–100.

[34] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *EUROSPEECH*, Geneva, Switzerland, 2003, pp. 2117–2120.

[35] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," National Institute of Standards and Technology (NIST), Gaithersburg MD, Tech. Rep.