The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Relative Phase Shift Features for Replay Spoof Detection System

Kantheti Srinivas[1], *Hemant A. Patil*[2]

Speech Research Lab,
Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India
{srinivas_kantheti, hemant_patil}@daiict.ac.in

Abstract

The replay spoofing tries to fool the Automatic Speaker Verification (ASV) system by the recordings of a genuine utterance. Most of the studies have used magnitude-based features and ignored phase-based features for replay detection. However, the phase-based features also affected due to the environmental characteristics during recording. Hence, the phase-based features, such as parameterized Relative Phase Shift (RPS) and Modified Group Delay are used in this paper along with the baseline feature set, namely, Constant Q Cepstral Coefficients (CQCC) and Mel Frequency Cepstral Coefficients (MFCC). We found out that the score-level fusion of magnitude and phase-based features are giving better performance than the individual feature sets alone on the ASV Spoof 2017 Challenge version 2. In particular, the Equal Error Rate (EER) is 12.58 % on the evaluation set with the fusion of RPS and the CQCC feature sets using Gaussian Mixture Model (GMM) classifier.

Index Terms: ASV, replay spoofing, Relative Phase Shift, Modified Group Delay.

1. Introduction

The recent technological developments in biometric applications lead to use of Automatic Speaker Verification (ASV) system. The ASV system either accepts or reject the claimed identity of a speaker based on the speech signal [1, 2]. Now-a-days spoofing detection is one of the important research areas in the field of the ASV system. There are various kinds of spoofing attacks for ASV, namely, impersonation, replay, identical twins, speech synthesis (SS), and voice conversion (VC) [1]. The ASV Spoof 2015 Challenge is the first edition which was mainly focussed to distinguish the natural speech and the artificial speech which is produced from SS and VC spoofing attacks [3]. On the other hand, the ASV Spoof 2017 Challenge focusses particularly on replay spoofing and its countermeasures in unknown conditions [4]. The replay spoofing can be classified as close and far-field recording [5]. Due to the availability of low cost and high quality recording and playback devices causes to increase the vulnerability of the ASV system.

Various countermeasures were proposed at the frontend and backend to classify the genuine and replay speech. The spectral bitmaps were proposed for replay Spoofed Speech Detection (SSD) task in a text-dependent speaker verification system [6]. In [5], the average spectral bitmaps and cosine-kernel score techniques were proposed to find the low frequency contents of a signal to distinguish the genuine and replay speech signal. The replay SSD can be done by analyzing the speech utterances based on acoustics and copy detection algorithm was proposed [4, 7]. The Constant Q Cepstral Coefficients (CQCC) were proposed in [8] and Gaussian Mixture Model (GMM) classifier was used. The CQCC, Mel Frequency Cepstral Coefficients (MFCC), and pitch (fundamental frequency, F_0) feature sets were used at the frontend, the ReliefF and minimum redundancy and maximum relevance algorithms were used for feature selection and giving as input to Support Vector Machine (SVM) classifier at the backend [9]. The long-term spectral statistics features were given as input to the proposed binary classifier for realistic type of attacks, such as presentation attacks and physical access [10]. At the frontend, the deep learning techniques, such as Convolutional Neural Network (CNN) with max feature map activation function and recurrent neural network and stacking of both the neural networks and SVM classifier at the backend were used for replay SSD task [11]. The single frequency filtering approach was proposed to highlight the channel variations in recorded signal and Bi-directional Long Short Term Memory (BLSTM) was used at the backend [12]. The various deep learning techniques were implemented for replay SSD task in [13, 14]. The Variable length Teager Energy Separation Algorithm Instantaneous Frequency Cosine Coefficients (VESA-IFCC) was proposed to predict the importance of IF in each subband energy via ESA to predict the possible changes in envelope spectrum with the help of cepstral coefficients to characterize the natural and the replay speech signal [15]. The experimental analysis of various spectral magnitude-based features and their score-level fusion using logistic regression were studied in [16]. The experiments were conducted on two databases, namely, ASV Spoof 2017 and AV Spoof [16]. In [17], the various experimental outcomes proved that the high frequency features can be used for replay SSD task. The speech-specific features were proposed, such as Glottal Closure Instants (GCIs), epoch features, peak-to-side-lobe ratio of mean and skewness and cepstral features and the combination of all feature sets is used for replay SSD task [18].

Less work has been done using phase-based features for replay attack detection. However, the phase contains significant information which may not be present in spectral magnitude-based features. The overview of various phase-based features and their applications is presented in [19]. In this study, the various phase-based features, such as Relative Phase Shift (RPS) [20, 21, 22] and Modified Group Delay Cepstral Coefficients (MGDCC) [23] are proposed to use for replay SSD task. The spectral magnitude-based features are also chosen to perform the score-level fusion with the phase-based features. The magnitude-based spectral features such as CQCC [24] and MFCC [25] are used.

2. Features Used

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

RPS is obtained by the transformation of instantaneous phases of the speech signal with the help of harmonic analysis [20]. The harmonic analysis considers the speech signal as the sum of periodic and aperiodic components and models these two com-

10.21437/SLTU.2018-21

ponents separately [26]. The periodic components are represented with the help of sinusoidal modeling [27] as the sum of sinusoidal harmonic components with the amplitudes, phases, and harmonic frequencies with an integer multiple of pitch frequency F_0 in each voiced frame of a speech signal. These harmonic amplitudes contains basic perceptual information of the signal [20]:

$$x(t) = \sum_{k=1}^{M} a_k sin(\phi_k(t)), \quad (1)$$

$$\phi_k(t) = 2\pi k F_0 t + \theta_k, \quad (2)$$

where M is the number of harmonic components, $\phi_k(t)$ and $\theta_k(t)$ are the instantaneous phase and initial phase shift of k^{th} harmonic components at any time instant, respectively. The instantaneous phase shift depends on the harmonic frequency kF_0 and the time instant, t [20]. On the other hand, the θ is constant if the waveform shape is stable under the local stationary condition which is independent of t [20]. From the above discussion, it is observed that the phase data represents two properties in the signal, namely, waveform shape and time synchronicity. The waveform shape depends upon the initial phase shift differences which is called as RPS [20]. The RPS is constant at a particular time instant t_p (i.e., $\phi_1(t_p) = 0$) and the RPS is shown at any time instant in Eq. (3) [20]:

$$r_k(t) = \phi_k(t_p) = \phi_k(t) - k\phi_1(t), \quad (3)$$

where $\phi_1(t)$ and $r_k(t)$ is the instantaneous phase of fundamental harmonic component and RPS at any t, respectively. The linear phase term ($2\pi k F_0 t$) which produces constant phase wrapping in instantaneous phase is not present in RPS [28]. It is the main advantage of RPS feature set. The RPS feature contains random phase values which are wrapped between $[-\pi, \pi]$ and zero values (cosine features) [22]. The Figure 1 shows the processing of RPS features for replay SSD task. In the processing of RPS feature, the Mel triangular filterbank is used instead of linear triangular filterbank [21, 29].

Figure 1: *Functional block diagram of RPS feature set.*

The first step in this process is an unwrapping operation. This operation is used to remove the wrapping discontinuities in each voiced frame of a speech signal. Whereas, the unwrapping operation generates the ambigious data different from the RPS values. Hence, the differentiation operation is applied to the unwrapped RPS data. Thereafter, the subband processing have done to the difference of unwrapped RPS values with the help of linear triangular filterbank and in addition, the average of differentiated unwrapped RPS values are appended with the filtered unwrapped RPS data. Furthermore, Discrete Cosine Transform (DCT) and Cepstral- Mean Normalization (CMN) is applied to obtain DCT-Linear-RPS feature vector.

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

The process of MGD feature extraction from the speech signal $(x(n))$ is the $x(n)$ is analyzed with the help of Short-Time Fourier Transform (STFT). The STFT of $x(n)$ is represented in magnitude and phase form as [30]:

$$X(\omega) = |X(\omega)|e^{j\phi(\omega)}. \quad (4)$$

The Group Delay (GD) function is used to extract the phase information ($\phi(\omega)$) from the speech signal. The GD is defined as the negative derivative of the FT phase w.r.t. frequency ω as shown in Eq. (5) [30]:

$$\tau(\omega) = -\frac{d}{d\omega}\phi(\omega) \quad = -imag[\frac{d}{d\omega}log(X(\omega))]. \quad (5)$$

The Eq. (5) can also be represented with the help of FT property as shown in Eq. (6) [23, 30]:

$$\tau(\omega) = \frac{X_r(\omega)Y_r(\omega) + X_i(\omega)Y_i(\omega)}{|X(\omega)|^2}, \quad (6)$$

where $Y(\omega)$ represents the FT of $nx(n)$, r is real part and i is imaginary part, respectively. The MGD was proposed to overcome the problems, namely, spikes and pitch periodicity effects in GD function [30]. The spikes problem can be reduced by cepstral smoothing of denominator term $|X(\omega)|$. The tuning parameters (ρ, γ) were introduced to reduce the spikes problems because it was not possible to reduce alone by cepstral smoothing process [30]. These parameters were tuned to which depends on that application. The MGD function is shown with tuning parameters as [31]:

$$\tau(\omega) = \frac{X_r(\omega)Y_r(\omega) + X_i(\omega)Y_i(\omega)}{|X_c(\omega)|^{2\rho}}, \quad \tau_m(\omega) = \frac{\tau(\omega)}{|\tau(\omega)|}|\tau(\omega)|^{\gamma}, \quad (7)$$

where $X_c(\omega)|$ is the smooth cepstral value of $|X(\omega)|$ and the values of tuning parameters vary between 0 to 1 [30]. The Algorithm of Mel frequency MGDCC (MFMGDCC) was explained in [32]. The Linear Frequency MGDCC (LFMGDCC) is obtained by the subband processing in MGD with the help of a linear triangular filterbank.

3. Experimental Results

3.1. Database and Classifier

The experimental results for replay SSD were performed on ASV Spoof 2017 Challenge version 2 database. The sampling rate of speech signal is 16 kHz with 16-bits resolution per sample. The details of database are given in Table 1. The data collection, partitions and number of speakers in this database is same as that of version 1 of ASV Spoof 2017 database.

Table 1: *Details of ASV Spoof 2017 Challenge Version 2 Database. After [33]*

Subset	Utterances	
	Genuine	Replay
Training	1507	1507
Development	760	950
Evaluation	1298	12008

The detailed explanation of version 2 and also the changes done in this database is explained in [33]. At the backend, the Gaussian Mixture Model (GMM) classifier is used. The GMM classifier consists of weighted Gaussian components based on three parameters, namely, mean, variance, and weights [34]. The speaker modeling is done by GMM based on acoustic variations using the training subset for genuine and replay speech

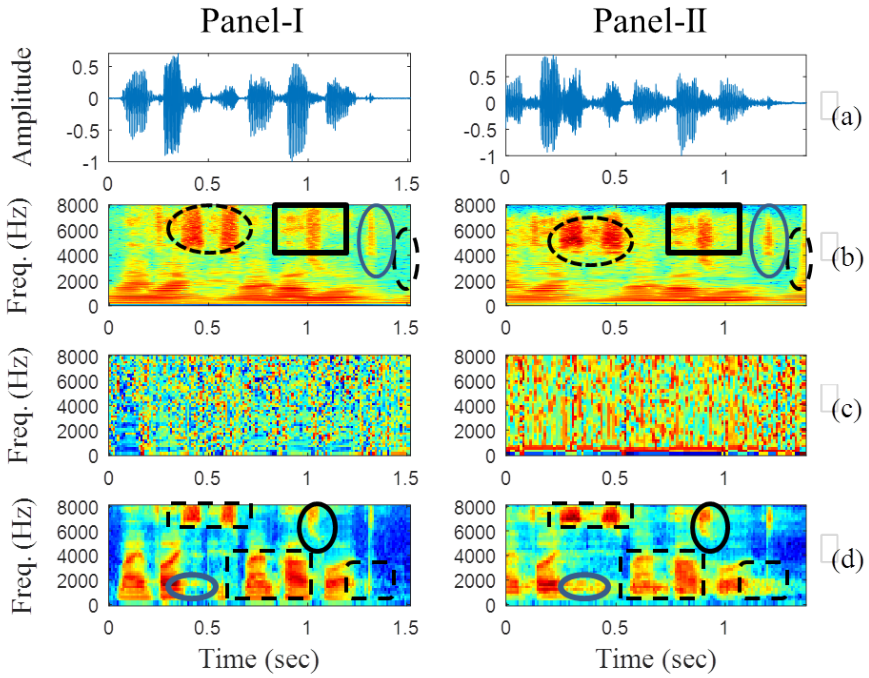

Figure 2: *(a) Speech utterance, (b) CQCC spectrogram, (c) linear filterbank unwrapped RPS phasegram, (d) MFCC spectrogram. Panel I and Panel II are for natural and replay signals, respectively.*

signals [35]. The development subset is used to optimize the threshold, and also assess the performance of the system. The purpose of testing the system with evaluation set is to assess the performance of the system in unknown conditions [4]. The scores obtained from the output of a classifier are the Log-Likelihood (LLK) scores. The score-level fusion is used to fuse the LLK scores to capture the possible complementary information in the system. The CQCC feature set contains 9 octaves and 96 number of bins per octave. The minimum and maximum frequency is 15 Hz and 8 kHz. The first 30 coefficients are retained using DCT and Δ, $\Delta\Delta$ coefficients are appended to form *90*-dimensional (D) feature vector and 512 number of GMM components in GMM is used, is the baseline system. The MFCC is extracted by the Hamming window of 20 ms duration and 10 ms shift and 40 number of Mel triangular filters in filterbank for subband processing results into *39*-D feature vector.
The DCT-Linear RPS feature set is extracted from the pre-processed (Pre-emphasis) speech signal. The rectangular window and 48 number of linear triangular filters in filterbank is used. The DCT is used for truncation to obtain first 13 coefficients that are appended along with Δ, $\Delta\Delta$ to form *39*-D feature vector, and Cepstral Mean Normalization (CMN) is applied for channel compensation, and the 128 number of GMM components are used.

Table 2: *Abbrevations Used for Feature Sets*

Feature Set	SSD system
CQCC (baseline)	Magnitude feature1 (M1)
MFCC	Magnitude feature2 (M2)
DCT-Linear-RPS	Phase feature1 (Ph1)
LFMGDCC	Phase feature2 (Ph2)

The LFMGDCC feature set is extracted from the pre-processed speech signals. The Blackman window is used for segmentation with 25 *ms* window duration, 10 *ms* window shift, and the 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 $\Delta\Delta$ coefficients are appended to form *39*-D feature vector and 512 number of GMM components are used. The Table 2 shows the various magnitude-based and phase-based features that are used for replay detection. The M1 feature set has high frequency resolution at low frequencies and high temporal resolution at high frequencies (i.e., the Q is constant across the entire frequency range) [8]. The Constant Q Transform (CQT) (which is extensively used in music signal analysis), uniform resampling and traditional cepstral analysis are used to obtain the M1 feature set [8]. In M2 feature set, the frequency resolution is linear up to 1 kHz and logarithmic after this value [25]. This feature set is obtained by the Short-Time Fourier Transform (STFT) and Mel triangular filterbank [25].
Figure 2 shows the analysis of a Ph1 feature set w.r.t. the spectral magnitude features, namely, M1, M2 features sets. Figure 2 shows that M1 feature set presents at low frequencies across the entire duration in both genuine and replay speech signals. The M1 spectrograms show some differences mainly at starting instants in genuine and the replay speech signals. The oval and rectangle shapes in M1 feature set are representing the intensity differences, the replay speech has better intensity than the genuine speech. The dotted oval at the end of utterance in spectrogram represents that the energy is present in replay speech whereas at that instat no energy present in genuine speech signal in M1 feature set. The phasegram of Ph1 feature represents the waveform shape which changes continuously in the speech signal. The phasegram shows significant differences in genuine and replay speech signals because this feature is sensitive to the noise. The M2 feature set captures the vocal tract information, these spectrograms of both the speech signals are bit

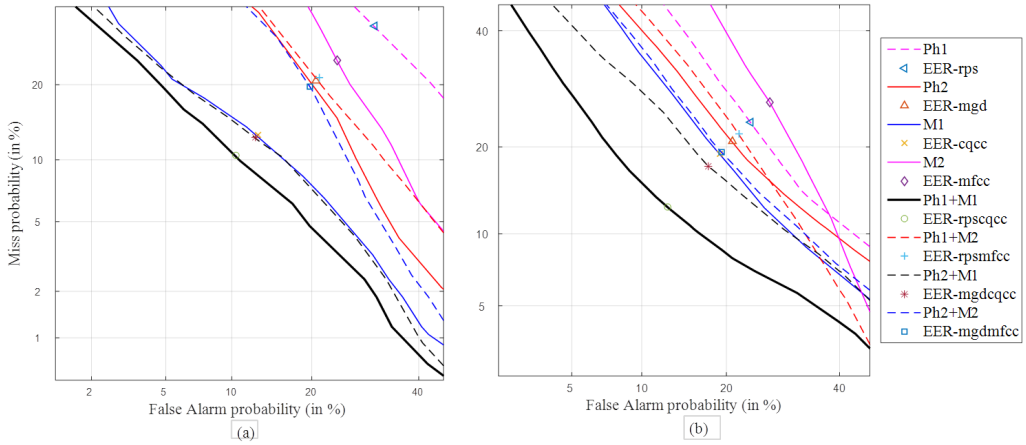

Figure 3: *DET curves of various replay SSD systems on (a) development set, and (b) evaluation set.*

looking different. In M2 feature set, the solid ovals and dotted rectangles are there one at the starting and the other at end position at the low frequencies indicates that the formants are absent in genuine whereas, the formants are present in replay speech. The frequency resolution in M2 feature set is less compared to the M1 feature set because of using STFT in M1 feature set. Hence, the strong complementary information lies in Ph1 fusion with M1. On the other hand, the complementary information is found to be less in Ph1 fusion with M2 feature sets.

Table 3: *Results (in % EER) of development and evaluation set*

SSD System	Dev	Eval
M1 (Baseline)	**12.81**	**19.04**
M2	24.19	26.90
Ph1	30.96	23.67
Ph2	20.70	20.84

The Table 3 shows the performance measures of various spectral magnitude-based, and phase-based features. It is observed that the Ph2 feature set is performing better than the M2 feature set in both development and evaluation datasets whereas the Ph1 feature set is performing better than the M2 feature set in evaluation dataset.

Table 4: *Results (in % EER) of score-level fusion of feature set*

Feature Set	Dev	Eval
M1 + Ph1	**10.41**	**12.58**
M1 + Ph2	12.45	17.43
M2 + Ph1	21.16	21.95
M2 + Ph2	19.73	19.28

Table 4 shows the results of the score-level fusion of spectral magnitude-based, and phase-based features. The score-level fusion of M1 with Ph1 and Ph2 the results are improved by 20.55 %, 11.09 % and 8.25 %, 3.41 % on development and evaluation dataset, respectively. Similarly, the fusion of M2 with Ph1 and Ph2 the results are improved by 9.80 %, 1.72 % and 0.97 %, 1.56 % on development and evaluation datasets, respectively. The better result is obtained with the fusion of M1 and Ph1 features set than the fusion of M2 and Ph1. Hence, the complementary information in Ph1 with M1 is found to be sufficient for replay SSD task. The Ph2 feature set contains less complementary information even its individual performance is better than the Ph1 feature set. The Ph2 feature set contains both magnitude and phase information. Hence, the performance is relatively poor for fusion with M1. However, the Ph2 feature set performance is better in fusion with M2 than with Ph1.

The Detection Error Tradeoff (DET) curve shown in Figure 3 indicates the performance curves at various operating points of ASV system [34] performance measures of the individual spectral magnitude-based and phase-based and the score-level fusion of feature sets. It is observed that the fusion of Ph1 and M1 is performing better than the individual magnitude-based and phase-based feature sets and other fused feature sets. The performance curve for fusion of the Ph1 and M1 is significantly improved in the evaluation set than the development set w.r.t. other feature sets.

4. Summary and Conclusions

This study showed the significance of various phase-based features in replay SSD task. The phase-based features, such as DCT-Linear-RPS and LFMGDCC features along with spectral magnitude-based features, namely, CQCC, MFCC feature sets are used in this study. Furthermore, these features are fused with each other. The CQCC feature set exhibits better performance than the other individual feature sets. The DCT-Linear-RPS feature provides complementary information better than the LFMGDCC feature with the fusion of CQCC feature set, even the LFMGDCC is giving better performance than the DCT-Linear-RPS feature set. The MFCC feature set gives better performance than the DCT-Linear-RPS feature set. The score-level fusion of DCT-Linear RPS feature set with the MFCC is not giving the better performance, indicating that the fusion of DCT-Linear-RPS with only CQCC gives the best performance. The LFMGDCC feature set is unable to give better performance than the DCT-Linear-RPS with the fusion of CQCC. Hence, the fusion of CQCC with only DCT-Linear-RPS is giving the best performance. We will explore the performance of these phase-based features by various neural network classifiers, such as Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) and Bi-directional LSTM.

5. 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] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification." in *INTERSPEECH*, Lyon, France, 2013, pp. 925–929.

[3] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASV Spoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2037–2041.

[4] 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.

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

[6] 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.

[7] 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.

[8] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop*, Bilbao, Spain, 2016, pp. 249–252.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[27] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[28] 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.

[29] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, "The AHOLAB RPS SSD spoofing challenge 2015 submission," in *Sixteenth Annual Conference of the International Speech Communication Association, INTERSPEECH*, Dresden, Germany, 2015, pp. 2042–2046.

[30] 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.

[31] 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.

[32] 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.

[33] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "Asvspoof 2017 version 2.0: metadata analysis and baseline enhancements," in *The Speaker and Language Recognition Workshop, Proc. Odyssey*, 2018, pp. 296–303.

[34] 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.

[35] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.