Novel Empirical Mode Decomposition Cepstral Features for Replay Spoof Detection

Conference Paper · August 2018

DOI: 10.21437/Interspeech.2018-1661

2 authors:

Prasad Anil Tapkir
Dhirubhai Ambani Institute of Information and Communication Technology
4 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Hemant Patil
Dhirubhai Ambani Institute of Information and Communication Technology
182 PUBLICATIONS **598** CITATIONS
SEE PROFILE

Interspeech 2018
2-6 September 2018, Hyderabad

Novel Empirical Mode Decomposition Cepstral Features for Replay Spoof Detection

Prasad A. Tapkir and Hemant A. Patil

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

Abstract

The advances in Automatic Speaker Verification (ASV) system for voice biometric purpose comes with the danger of spoofing attacks. The replay attack is the most accessible attack, where the attacker imitates speaker's identity by replaying the pre-recorded speech samples of the target speaker. Most of the conventional features, such as Mel Frequency Cepstral Coefficients (MFCC), Instantaneous Frequency Cepstral Coefficients (IFCC), etc. uses filterbank structure for feature extraction purpose. In this paper, we propose a novel Empirical Mode Decomposition Cepstral Coefficient (EMDCC) feature set, where the filterbank in MFCC is replaced with the Empirical Mode Decomposition (EMD) to obtain the subband signals. The proposed feature set takes an advantage of using EMD that acts as a dyadic filterbank and handles the nonlinear and non-stationary nature of the speech signal. The stand-alone EMDCC feature set gives the Equal Error Rate (EER) of 28.06 % compared to the baseline CQCC and MFCC system with EER of 29.18 % and 31.3 %, respectively on the evaluation set of ASV Spoof 2017 Challenge database. Furthermore, the proposed feature set is fused with the Linear Frequency Modified Group Delay Cepstral Coefficient (LFMGDCC) at score-level and we obtain a reduced EER of 18.36 % on evaluation set.

Index Terms: replay, spoofing, empirical mode decomposition, dyadic filterbank, residual.

1. Introduction

In the past few years, there has been noteworthy hike in the use of Automatic Speaker Verification (ASV) system for numerous applications, such as security, telephone banking etc. [1]. In practice, the ASV system should be robust across various variabilities, such as speaker aging, microphone, transmission channel etc. Nullifying the effects of variabilities of these components makes the ASV system robust, however, it comes with the disadvantage of making the ASV system vulnerable to spoofing attacks. Hence, it is necessary to make the ASV system secure against spoofing attacks. The various kinds of spoofing attack include impersonation [2], Voice Conversion (VC) [3], Speech Synthesis (SS) [4] and replay [5]. Replay is the simplest and easily accessible spoofing attack, as it does not require any prior knowledge of specific expertise and special computer knowledge [6]. Replay is a spoofing attack, where the attacker tries to fool ASV system using a pre-recorded speech samples of the target genuine speaker [6].

The replay speech can be mathematically model as a convolution of the genuine speech signal with the impulse response of recording device, the impulse response of recording environment, the impulse response of multimedia speaker (playback device) and impulse response of playback environment [5]. The quality of recording/ playback device, and noise level in recording/ playback environment decides the quality of replay speech and hence the difficulty of replay spoof detection. The replay speech recorded with a high quality recording and playback device in clean recording environment is very difficult to detect as it is very much similar to the genuine speech.

One of the initial study in replay detection for text-dependent ASV used score normalization approach and decision was made based on N-similarity scores [7]. The channel noise from recording device was used to detect the replay speech [8]. The replay speech obtained through far-field recording using land line and GSM telephone channel using modulation index and spectral ratio was studied in [9] for text-dependent ASV system and in [10] for text-independent ASV system. Recently the second ASV spoof 2017 challenge is organized to develop a countermeasure to prevent replay attack under unseen conditions [11]. The high frequency gets highly affected and hence most of the cues for replay spoof detection can be found in the high frequency region [12]. Instantaneous Frequency (IF)-based approach was studied in [13]. Single Frequency Filtering (SFF) approach was used to capture channel information present along with generative and discriminative model using Gaussian Mixture Model (GMM) and Bidirectional Long Short Term Memory (BLSTM) classifier at the back end [14]. The source-based features, namely, Epoch Feature (EF) and Peak to Side lobe Ratio Mean and Skew (PSRMS) were used in [15]. The score-level fusion of these feature with IFCC, Mel Frequency Cepstral Coefficients (MFCC) and Constant Q Cepstral Coefficients (CQCC) was carried out to capture complementary information. To detect known and unknown replay audio, effective ensemble learning classifier was proposed along with various acoustic features in [16]. Various neural network-based countermeasures were developed in [17, 18, 19, 20].

Several conventional features, such as MFCC, LFCC, IFCC, etc. uses filterbank structure to obtain the subband filtered signal for processing. This filterbank is fixed for all the utterances in feature extraction process. In addition, these feature extraction methods assume that the input speech signal is a stationary signal and produced by linear system. The Empirical Mode Decomposition (EMD) were first used for non linear and non-stationary time-series analysis [21]. The EMD decomposes signal into special Intrinsic Mode Functions (IMFs), assuming that the input signal is produced by non linear system and is non-stationary in nature. In addition, all the events are handled as they arise. Flandrin *et. al.* reported that the EMD acts as a dyadic filterbank and decomposes input signal similar to wavelet-like decomposition [22]. In this paper, we propose a new feature extraction approach in which the filterbank is replaced with EMD to capture all the advantages of EMD.

10.21437/Interspeech.2018-1661

We refer to this feature set as Empirical Mode Decomposition Cepstral Coefficients (EMDCC). The experimental results are compared with MFCC [23] and CQCC [24] results. Furthermore, the EMDCC feature set is fused with the LFMGDCC at score-level to capture possible complementary information.

2. Empirical Mode Decomposition (EMD)

The EMD deals with the signals that are produced by non linear system and non-stationary in nature and the events are handled as the signal occur [25]. EMD decomposes a signal into zero mean Amplitude Modulation and Frequency Modulation (AM-FM) waveforms. These AM-FM waveforms are known as Intrinsic Mode Functions (IMFs), and aims to represent underlying intra-wave modulated components in the signal. The *locally* zero mean condition is assured by maintaining mean of the lower and upper envelope *(obtained by interpolating minimas and maximas)* of an IMF equal to zero [21, 26]. For a given signal $s(n)$, it can be decomposed into two part, one is high-frequency (local) part also known as *detail*, $d(n)$ and the other part is low-frequency (local) part known as *trend*, $m(n)$ [22].

$$s(n) = d(n) + m(n). \tag{1}$$

To obtain the further *detail* and *trend* signal, the procedure of decomposing signal into *detail* and *trend*, iteratively applied on residual term considering as a new signal. For given signal $s(n)$, the IMFs can be found as follows [21]:

1. Identify all the relative minima and maxima of the signal.
2. Interpolate maxima (and minima) to get upper envelope $e_{up}(n)$ (and lower envelope $e_{low}(n)$).
3. The mean of upper and lower envelope is computed as,

$$m_{env}(n) = \frac{e_{up}(n) + e_{low}(n)}{2}. \tag{2}$$

4. To get rid of overriding wave, the mean signal $m_{env}(n)$ is subtracted from the original signal $s(n)$.

$$x_1(n) = s(n) - m_{env}(n). \tag{3}$$

   Due to small humps present in speech signal and which are not identified as minima or maxima in the first step, $x_1(n)$ cannot be said as IMF. Hence, to get the actual IMF the process continued iteratively, considering $x_1(n)$ as new signal, sifting process is continued to obtained $x_{11}(n)$. The following series is called $x_{1j}(n)$. The procedure is repeated till the following condition is meet by the residue:

$$\sum_{n=0}^{N} \left[\frac{|x_{1(j-1)}(n) - x_{1j}(n)|^2}{(x_{1(j-1)}(n))^2} \right] \leq 0.2. \tag{4}$$

   Assuming $d_1(n) = h_{1J}$, satisfies condition in Eq. (4), $d_1(n)$ is called as first IMF.

5. Subtract the IMF $d_1(n)$ from the actual signal $s(n)$,

$$m_1(n) = s(n) - d_1(n). \tag{5}$$

6. Considering this $m_1(n)$ as a new signal, the entire process (step 1- step 5) is repeated to obtain other IMFs. For stoping either of the following two conditions should be satisfied.
   - The signal does not contain any minima and maxima i.e., signal becomes void or monotonic.
   - The signal level becomes negligible such that it can be neglected for further processing.

Finally, for a given signal $s(n)$, having I number of IMFs the EMD can be represented as,

$$s(n) = m_I(n) + \sum_{i=1}^{I-1} d_i(n), \tag{6}$$

where $m_I(n)$ represents residual term and $d_i(n)$, $(i = 1, 2, ..., I-1)$ represents zero mean AM-FM waveforms (IMFs/modes).

Figure 1 shows the IMFs obtained for the voiced frame of the speech signal. The first IMF obtained is high frequency signal indicating the filter for IMF: 1 or mode: 1 is high pass in nature, while the other filters associated with another IMFs or modes are bandpass in nature [22]. The narrowband condition for decomposed signals can be assured by maintaining number of zero crossings of components to be either same or differ by at most one with number of minima and maxima.

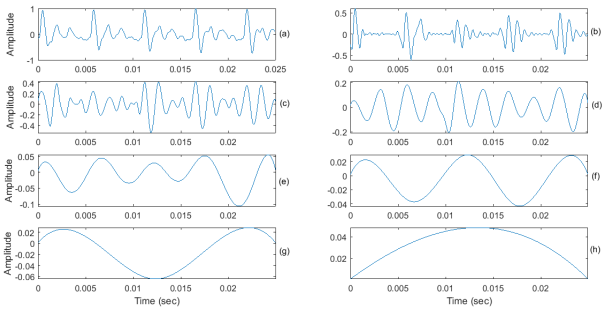

Figure 1: *(a) Voiced frame of speech signal, (b-g) IMF: 1-6 and (h) IMF: 7 or trend.*

The EMD behaves as a dyadic filterbank and decomposes signal similar to the wavelet like decomposition for fractional Gaussian noise [22] and for white noise [27]. From the process of EMD, each IMF is zero mean AM-FM waveform whose number of zero crossings is equal or differ by at most one with its number of extremas. The number of zero crossings can be roughly interpreted as mean frequency of each IMF. Empirical results shows that the number of zero crossing resemble to log_2 scale and shows equivalent structure of filterbank with log scale [22].

Figure 2(a) shows the spectrogram of first 11 IMFs obtained by EMD decomposition and Figure 2(b) shows the spectrogram of first 11 subbands obtained by passing speech signal through mel triangular filterbank of 11 filters. From Figure 2, it is observed that the IMF spectrogram is very much similar to spectrogram of subband obtained through filterbank, however, EMD gives the very good resolution at the lower frequencies and poor resolution at high frequencies compared to filterbank *(due to log_2 scale)*. It is also observed that EMD behaves as a very sharp filter blocking for most of the frequencies outside the range of IMF frequencies compared to filterbank, which passes other frequencies with less attenuation.

3. Feature Extraction

3.1. Proposed EMDCC Feature

Figure 3 shows the functional block diagram to extract the proposed EMDCC feature set. The input speech signal is first decomposed into I zero-mean AM-FM waveforms (IMFs). Each

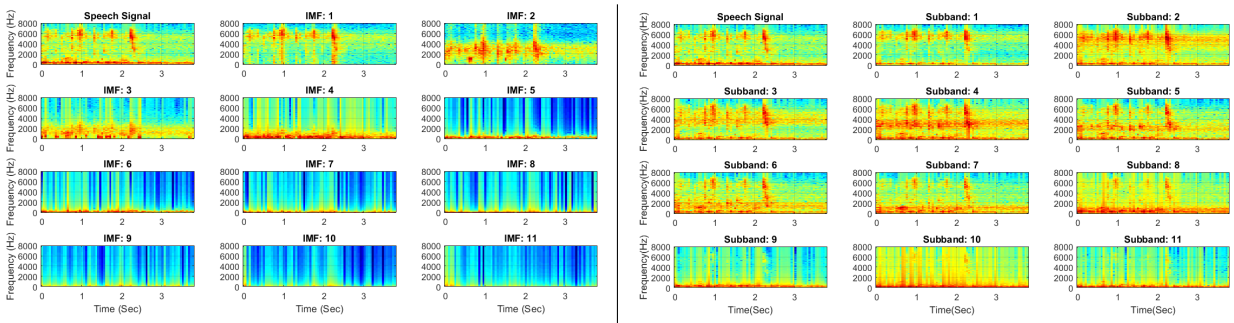

Figure 2: *Spectrogram of first 11 IMFs/subbands obtained through (a) EMD decomposition and (b) Mel triangular filterbank.*

IMF is then segmented with 20 *ms* duration with 50 % overlap. Further, the energy in each frame $x(n)$ is computed for every IMF as follows:

$$E = \sum_{m=0}^{M-1} |x(m)|^2, \tag{7}$$

where M is the length of frame (320 samples) and E represents energy in frame $x(n)$. To approximate the non linear relation between auditory-nerve firing rate and signal intensity, log function is applied on the computed energies. This nonlinearity provides remarkable robustness by suppressing small signal variability and also mimics the human perception of loudness [26, 28].

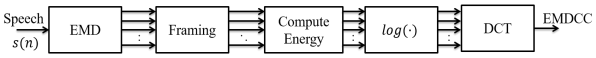

Figure 3: *Functional block diagram to extract proposed EMDCC feature set.*

The Discrete Cosine Transform (DCT) is used to exploit the redundancy and to obtain the cepstral coefficients. First few cepstral coefficients are retained and in order to capture the dynamic nature of the speech signal they are appended with their first and second order derivatives to obtain high dimensional feature vector.

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

The phase spectrum need complex unwrapping algorithm before processing, hence the use of phase spectrum in speech applications has been ignored for several years. In this paper, to capture the phase information LFMGDCC feature set is used with EMDCC feature set. The Group Delay (GD) function which is defined as negative frequency derivative of phase function, has same properties as the phase function [29]. Speech signal is a output of a stable system, hence only zeros (due to noise or analysis window) are very close to unit circle. These zeros causes the GD function to be spiky in nature. The Modified Group Delay (MGD) function, proposed in [30] suppresses these zeros. The GD function $G(\omega)$ for speech frame $s(n)$ can be computed as [31]:

$$G(\omega) = \frac{S_r(\omega)Q_r(\omega) + S_i(\omega)Q_i(\omega)}{|S(\omega)|^2}, \tag{8}$$

where $S(\omega)$ and $Q(\omega)$ represents Fourier Transform (FT) of $s(n)$ and $ns(n)$ respectively (suffix r and i indicates real and imaginary part respectively). To suppress the zeros close to unit circle, the denominator term is replaced by cepstrally smoothed spectra ($R(\omega)$) of $S(\omega)$.

$$G(\omega) = \frac{S_r(\omega)Q_r(\omega) + S_i(\omega)Q_i(\omega)}{|R(\omega)|^{2\rho}}, \tag{9}$$

$$G_m(\omega) = \frac{G(\omega)}{|G(\omega)|}(|G(\omega)|^{\gamma}), \tag{10}$$

where $G_m(\omega)$ represents MGD function. The parameters ρ and γ decides the reduction level in amplitude of spikes and restores the dynamic range in GD function and needs to be tuned as per application. To use this MGD fuction for replay spoof detection, the cepstral coefficients are computed similar to Mel Frequency Modified Group Delay Cepstral Coefficients (MFMGDCC) [32] except that the Mel scale is replaced with linear scale. We refer to this feature set as LFMGDCC. The extraction of LFMGDCC start with computing MGD values and passing them through linear triangular filterbank. The energy in each filter of the filterbank is computed, and to decorrelate the signal DCT is applied [33]. First few coefficients are retained and are appended with Δ and $\Delta\Delta$ coefficients to capture dynamic information.

4. Experimental Setup and Results

4.1. Database

All the experiments are performed on ASV spoof 2017 challenge database. The database is based on the text-dependent RedDots corpus and its replayed version. All the speech signals have sampling frequency of 16 kHz and 16 bit per sample precision. The database contains three subsets, namely, training, development and evaluation set. The details of database are given in [11]. The GMM classifier is used to classify between genuine and replay speech. Two GMMs are trained for genuine and replay speech using training set of ASV Spoof 2017 Challenge database.

4.2. EMDCC

The cubic spline interpolation is used to interpolate maxima and minima to obtain upper and lower envelope. The first 10 IMFs obtained from EMD are used for the feature extraction. The 10 IMFs are segmented into short frame of duration 20 *ms* with 50 % overlap. The 10 static coefficients are appended with Δ and $\Delta\Delta$ coefficients, resulting in *30*-dimensional (D) feature vector. The GMM classifier with 512 Gaussian components is used.

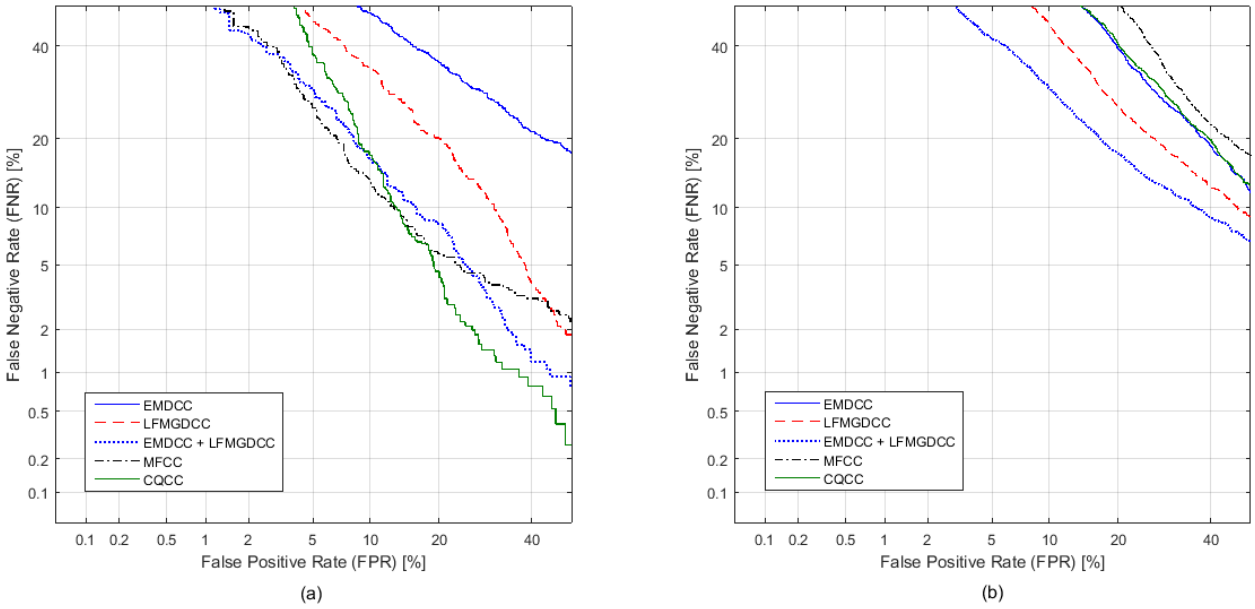

Figure 4: *DET curves for (a) development and (b) evaluation set.*

4.3. LFMGDCC

The parameter ρ and γ are set to *0.4* and *0.1* respectively for computing MGD function. The 13-D static LFMGDCCs along with Δ and $\Delta\Delta$ coefficients are used to get 39-D feature vector. The feature were extracted using 40 linearly scaled triangular filters along with the Hamming window of 25 ms duration and 10 ms shift. The GMM classifier with 512 Gaussian components is used.

4.4. Results

The results of our proposed feature set along with other feature sets are shown in Table 1. We have compared our proposed feature set with CQCC (baseline system) and MFCC feature set.

Table 1: *Results on development and evaluation data set*

Feature Set	**EER (%)**	
	Development	**Evaluation**
CQCC *(Baseline)*	**12.11**	29.18
MFCC	12.21	31.3
LFMGDCC	19.26	22.91
EMDCC	28.48	28.06
EMDCC+LFMGDCC	12.42	**18.36**

'+' indicates the score-level fusion.

The CQCC-GMM system is a baseline system provided by the organizers of the ASV Spoof 2017 Challenge, having EER of 29.18 % on evaluation set. The MFCC-GMM and LFMGDCC-GMM systems gives an EER of 31.3 % and 22.91 % on evaluation set. The EMDCC gives result of 28.06 % compared to MFCC system on evaluation set giving improvement of 3.24 %. Further, EMDCC feature set is fused with phase-based LFMGDCC feature at score-level giving reduced EER of 12.42 % and 18.36 % on development and evaluation set respectively, indicating that the feature sets captures complementary information. Figure 4 shows the DET curves [34] for the development and evaluation set of the ASV spoof 2017 challenge database. From DET plots, it is clear that score-level fusion of EMDCC and LFMGDCC gives better performance over entire operating region of DET curve. The authors also performed experiments with power law nonlinearity, where the exponent value $-\frac{1}{30}$ gave best results. Using power law nonlinearity, the EER of 28.99 % and 27.87 % on development and evaluation set of ASV spoof 2017 challenge database respectively is obtained. However, It is observed that log nonlinearity captures better complementary information compared to power law nonlinearity.

5. Summary and Conclusions

In this paper, we proposed novel EMDCC feature for replay spoof detection task. The EMD decomposes a signal into IMFs assuming that the signal is produced by non linear system and is non-stationary in nature. In addition, EMD act as dyadic filterbank. The proposed feature set uses EMD to decompose speech signal into subbands and uses log nonlinearity to approximate non linear relation between auditory nerve firing rate and signal intensity. Further, this feature set is fused with phase-based feature LFMGDCC at score-level, to capture complementary information. We also observed the individual EMDCC system performance for power law nonlinearity is better, however log nonlinearity capture better complementary information. The results of the final fused system are compared with baseline CQCC and MFCC system. The EMDCC-GMM system performs relatively better than baseline CQCC-GMM and MFCC-GMM system. In future, neural network-based classifiers can be used to improve the performance of spoof detection system. In addition, replay spoof detection system can be implemented using channel noise estimated by EMD.

6. Acknowledgement

The authors would like to thank Mr. Ankur Patil and Mr. Srinivas Kantheti for their valuable suggestions. The authors would also like to thank authorities of DA-IICT, Gandhinagar to carry out this research work.

7. 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] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *IEEE International Symposium on Intelligent Multimedia, Video and Speech Processing*, 2004, pp. 145–148.

[3] Y. Stylianou, "Voice transformation: A survey," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 3585–3588.

[4] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

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

[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 in Asia-Pacific Signal and Information Processing Association, (APSIPA)*, Chiang Mai, Thailand, 2014, pp. 1–5.

[7] W. Shang and M. Stevenson, "Score normalization in playback attack detection," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, USA, 2010, pp. 1678–1681.

[8] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *IEEE International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 4, 2011, pp. 1708–1713.

[9] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," *Biometrics and ID Management*, Brandenburg, Germany, pp. 274–285, 2011.

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

[11] 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.

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

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

[14] 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 2017," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 107–111.

[15] 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.

[16] Z. Ji, Z.-Y. Li, P. Li, M. An, S. Gao, D. Wu, and F. Zhao, "Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof 2017," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 87–91.

[17] 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.

[18] 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.

[19] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and model fusion for automatic spoofing detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 102–106.

[20] 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.

[21] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," in *The Royal Society of London A*, vol. 454, 1998, pp. 903–995.

[22] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 112–114, 2004.

[23] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *Readings in Speech Recognition*. Elsevier, 1990, pp. 65–74.

[24] 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*, vol. 25, 2016, pp. 249–252.

[25] N. E. Huang, *Hilbert-Huang transform and its applications*. World Scientific, 2014, vol. 16.

[26] X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney, "A phenomenological model for the responses of auditory-nerve fibers: I. nonlinear tuning with compression and suppression," *The Journal of the Acoustical Society of America (JASA)*, vol. 109, no. 2, pp. 648–670, 2001.

[27] Z. Wu and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method," in *The Royal Society of London A*, vol. 460, 2004, pp. 1597–1611.

[28] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *The Journal of the Acoustical Society of America (JASA)*, vol. 106, no. 4, pp. 2040–2050, 1999.

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

[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] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication*, vol. 81, pp. 54–71, 2016.

[32] 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)*, Vancouver, BC, Canada, 2013, pp. 7234–7238.

[33] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete Cosine Transform," *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[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, Gaithersburg MD, Tech. Rep., 1997.