IMPORTANCE OF ANALYTIC PHASE OF THE SPEECH SIGNAL FOR DETECTING REPLAY ATTACKS IN AUTOMATIC SPEAKER VERIFICATION SYSTEMS

Shaik Mohammad Rafi B and K Sri Rama Murty*

Department of Electrical Engineering, Indian Institute of Technology Hyderabad, India
ee17resch01003@iith.ac.in, ksrm@iith.ac.in

ABSTRACT

In this paper, the importance of analytic phase of the speech signal in automatic speaker verification systems is demonstrated in the context of replay spoof attacks. In order to accurately detect the replay spoof attacks, effective feature representations of speech signals are required to capture the distortion introduced due to the intermediate playback/recording devices, which is convolutive in nature. Since the convolutional distortion in time-domain translates to additive distortion in the phase-domain, we propose to use IFCC features extracted from the analytic phase of the speech signal. The IFCC features contain information from both clean speech and distortion components. The clean speech component has to be subtracted in order to highlight the distortion component introduced by the playback/recording devices. In this work, a dictionary learned from the IFCCs extracted from clean speech data is used to remove the clean speech component. The residual distortion component is used as a feature to build binary classifier for replay spoof detection. The proposed phase-based features delivered a 9% absolute improvement over the baseline system built using magnitude-based CQCC features.

Index Terms— Convolutional distortion, dictionary learning, Gaussian mixture models, IFCC features.

1. INTRODUCTION

Unlike artificial person authentication techniques such as passwords and identity cards, biometric techniques are going to be the future of person authentication. Voice is an important personal trait used in speaker authentication systems. Automatic speaker verification (ASV) system is a powerful biometrics solution as the verification could be done remotely through voice communication. Unfortunately, ASV systems are vulnerable to various spoofing attacks [1, 2]. Intentional circumvention in the security of ASV systems using fake audio recordings is referred to as spoofing attacks or presentation attacks. Spoofing attacks include impersonation, speech synthesis, voice conversion and replayed speech. Most spoof detection studies focus on the first three attacks and reliable counter measures have been developed. The studies on replay attack accelerated after the ASV spoof challenge [3].

In the light of recent events involving replay attacks, counter measures have been receiving much attention. Replay attacks are so simple that any one can easily record the genuine person's voice with high quality recording devices and it does not require any expertise in signal processing. Efforts are afoot to design spoof detection systems that discriminate fake signals from authentic ones [4]. The performance degradation of ASV systems in case of spoofing attacks was demonstrated by Villalba et al. [5]. In the recent past, researchers have been pursuing the problem from the aspects of feature extraction/representation and system level modelling for classification of the features.

Among the various kinds of speech features, magnitude spectral features extracted from the short-time Fourier transform (STFT) of the speech signal are more popular in speech applications. Massimiliano Todisco, et al., in their seminal work [6] introduced specialized features called Constant Q Cepstral Coefficients (CQCCs) for spoof detection. These features are extracted by applying real cepstral analysis on the constant Q transform (CQT) [7] of the speech signal. The placement of frequency bins of the CQT is motivated from the human perception mechanism [8]. In contrast to the linear frequency bins of the Fourier transform, the center-frequencies of the CQT filters are geometrically spaced in order to maintain a constant Q-factor, across the filters. This provides variable frequency-resolution, and captures detailed characteristics which are missed in other feature extraction techniques.

Recently, ASVspoof 2017 challenge that aims at the task of replay spoof detection, aided the researchers to explore the solution to the problem in different perspectives. The summary of the challenge with preliminary evaluation results can be found in [9]. An experimental study of different features on spoof detection systems was presented in [10]. Different steps in the development of spoof detection systems was investigated in [11].

When speech is replayed through a playback device, the acoustic characteristics have very subtle differences from the original speech, thus making it hard to detect the spoof attack. The distortion introduced by the recording/playback device for detecting the replay attacks [12]. In [13], an estimate of the live speech component is subtracted from the replayed speech, in the magnitude spectral domain, to highlight the device characteristics. In this paper, we explore the importance of analytic phase of the speech signal in replay spoof detection, using instantaneous frequency cosine coefficients (IFCC) features. The performance of the proposed IFCC features is evaluated on ASVspoof-2017 database and compared with constant-Q cepstral coefficients (CQCC), mel-frequency cepstral coefficients (MFCC) and modified group-delay coefficients (MDGC).

The rest of the paper is organized as follows. Section 2 discusses the IFCC extraction from analytic phase of speech signal. In Section 3, dictionary learning based method is described to highlight the device characteristics in IFCC features. Section 4 presents experimental evaluation of the proposed features, and compares them with the baseline results. Section 5 summarizes this work and points to some possible future directions.

*The first author is an employee of IIIT-RK Valley Campus, RGUKT-AP, Andhra Pradesh, India

2. FEATURE EXTRACTION FROM INSTANTANEOUS FREQUENCIES

The analytic signal $s_a(t)$ of a continuous-time signal $s(t)$ contains only the positive frequency components [14]. The analytic signal can be computed in the time-domain as

$$s_a(t) = s(t) + j s_h(t), \tag{1}$$

where $s_h(t)$ is the Hilbert transform of $s(t)$, given by

$$s_h(t) = \frac{1}{\pi t} * s(t). \tag{2}$$

The analytic signal in (1) can be expressed in the polar form as

$$s_a(t) = |s_a(t)| \exp(j\phi(t)) \tag{3}$$

where $|s_a(t)| = \sqrt{s^2(t) + s_h^2(t)}$ and $\phi(t) = \tan^{-1}\left(\frac{s_h(t)}{s(t)}\right)$ denote the instantaneous amplitude and instantaneous phase of $s(t)$, respectively. The time-derivative of the unwrapped instantaneous phase is referred to as the instantaneous frequency (IF), and is given by

$$\phi'(t) = \frac{d\phi(t)}{dt} \tag{4}$$

Even though IF can be computed for any arbitrary signal, it has a physical interpretation only when $s(t)$ is a narrowband (NB) signal. Hence, a wideband signal like speech is typically passed though a bank of K NB filters, and IF is computed on the filtered outputs. The analytic signal of the k^{th} filter output $s_k(t)$, centered at Ω_k, can be expressed as

$$s_{ak}(t) = |s_{ak}(t)| \ \exp(\Omega_k t + \theta_k(t)) \tag{5}$$

where $\theta_k(t)$ denotes the instantaneous deviation of the phase from $\Omega_k t$. For the NB case, the IF

$$\phi'_k(t) = \Omega_k + \theta'(t), \tag{6}$$

provides a measure for IF deviation $\theta'_k(t)$ of the signal $s_k(t)$ from its center frequency Ω_k [15].

The IF of a signal is a measure which is often of significant practical importance [15]. However, in practice, the computation of IF suffers from the phase wrapping problem [16]. In order to circumvent the phase wrapping problem, we use the IF computation method using Fourier transform relations, proposed in [17]. In this method, the IF of the k^{th} filtered component $s_k[n]$ can be computed in the discrete-domain as

$$\phi'_k[n] = \frac{2\pi}{N} \Re \left\{ \frac{\mathcal{F}^{-1}(l \ S_k[l])}{\mathcal{F}^{-1}(S_k[l])} \right\} \tag{7}$$

where N is the length of the NB signal $s_k[n]$, $S_k[l]$ is the $N-$point discrete Fourier transform of $s_k[n]$, $\Re$ denotes the real-part and $\mathcal{F}^{-1}$ denotes the inverse discrete Fourier transform.

NB components of the speech signal are extracted using a bank of K linearly spaced Gaussian shaped NB filters. IF is estimated for each of these K NB components. In order to estimate frame-level features, the IF contours are averaged over a short frames of 25 ms shifted by 10ms, resulting in a $K-$dimensional mean IF vector. Discrete cosine transform is applied on the $K-$dimensional mean IF vector to pack the information compactly, and first 20 dimensions are retained for modeling. The resulting lower dimensional representation extracted from the IF are referred to as the instantaneous frequency cosine coefficients or IFCC features.

3. HIGHLIGHTING DEVICE-SPECIFIC FEATURES USING DICTIONARY LEARNING

The aim of replay spoof detection is to determine that a given speech utterance is from a live speaker or an intermediate recording/playback device. The device can be assumed to be a linear time invariant (LTI) system with an impulse response of $h(t)$, the replayed signal $r(t)$ can be expressed in terms of the live speech signal $s(t)$ as,

$$r(t) = s(t) * h(t). \tag{8}$$

In the case of live-speech, the impulse response $h(t)$ reduces to $\delta(t)$, the ideal distortion less channel. Development of a robust spoof detection system requires features that highlight the intermediate device characteristics, i.e., we need to extract features that characterize $h(t)$ from the recorded signal $r(t)$. The convolutive relationship in (8) transforms to multiplicative relationship in the frequency domain, and is given by

$$R(j\Omega) = S(j\Omega)H(j\Omega). \tag{9}$$

As a consequence, the Fourier phase and the group-delay (GD) admit an additive relationship, i.e.,

$$\tau_r(\Omega) = \tau_s(\Omega) + \tau_h(\Omega) \tag{10}$$

where $\tau_r(\Omega)$, $\tau_s(\Omega)$ and $\tau_h(\Omega)$ are the group-delays GDs of $r(t)$, $s(t)$ and $h(t)$ respectively. While the additive relation of phases Fourier domain is exact, such a relation is not straightforward for phases in the analytic signal domain. For case of asymptotic signals, the class of signals whose IF and GD relations are approximately the same function of the whole range of frequencies, the relationship between IFs of the convolved signal can be derived from their GDs [18]. In this paper, we assume that the deviations of IF $\theta'_k(t)$ of $r(t)$, $s(t)$ and $h(t)$, from the center frequency Ω_k are related by

$$\theta'_{rk}(t) = \theta'_{sk}(t) + \theta'_{hk}(t). \tag{11}$$

Hence the IF of the replayed speech signal contains additive distortion introduced by the device characteristics.

Pyknograms in Fig. 1 illustrates the effect of distortion introduced by the device characteristics in the IF domain. Pyknogram visualizes the IF variations along the time-frequency plane Fig. 1(a) shows pyknogram of a live speech signal, while Fig. 1(b) shows pyknogram of its replayed version, and hence it suffers from device distortion. The additive distortion introduced by the intermediate playback/recording device disturbs the density of IF contours, leading to distortion of formant structure. This phenomenon can be observed in the high frequency region of replayed signal in Fig. 1(b).

The IFCC feature vectors extracted from the replayed speech signal $\mathbf{r}$ contains an additive combination of clean speech component $\mathbf{s}$ and device distortion component $\mathbf{h}$, i.e., $\mathbf{r} = \mathbf{s} + \mathbf{h}$. In the case of live speech recording, the device distortion $\mathbf{h}$ should be ideally zero. Hence, in order to distinguish the live speech from replayed speech, we need to rely on the features highlighting the device distortion component $\mathbf{h}$. Since, we do not have direct access to the device distortion component $\mathbf{h}$, we propose to subtract an estimate of the live speech component $\hat{\mathbf{s}}$ from the replay speech $\mathbf{r}$ to highlight the device distortion component $\mathbf{h}$. In our earlier work using magnitude spectral features, we proposed a dictionary learning based method for this purpose [13].

An estimate of the live speech component $\mathbf{s}$ in the replayed speech $\mathbf{r}$ can be obtained by it onto the acoustic space spanned by the live speech. The acoustic space spanned by live speech is modeled

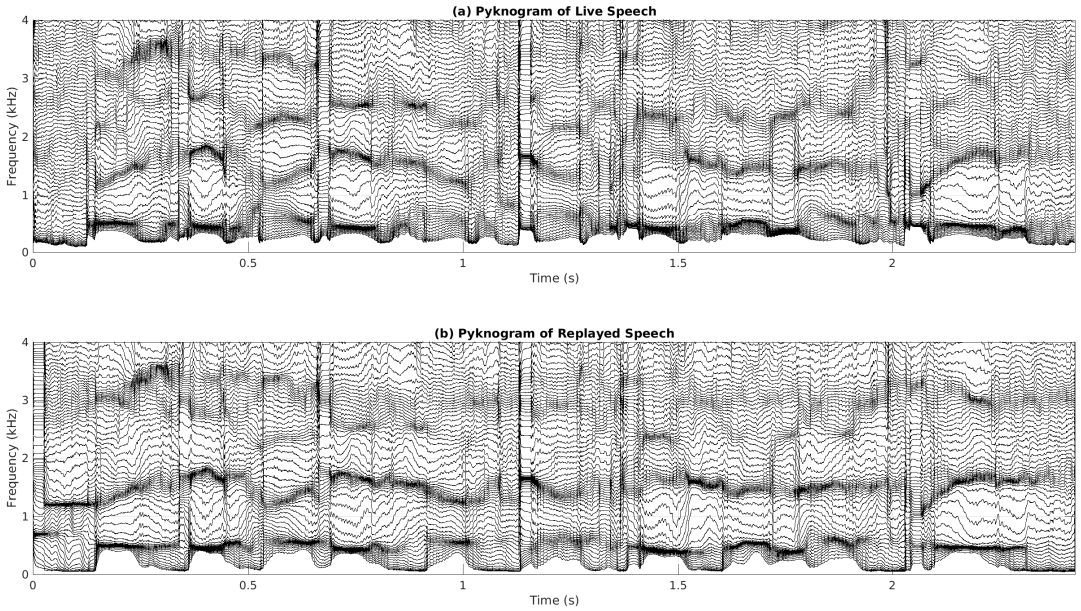

Fig. 1. Effect of device characteristics on IF. (a) Pyknogram of live speech and (b)Pyknogram of replayed speech.

by learning an overcomplete dictionary $\mathbf{A}$ from the live speech data. Sparse representations have been shown to be effective for speaker identification [19]. The K-singular value decomposition (K-SVD) algorithm, proposed by Aharon et al., is used for joint optimization of dictionary atoms and sparse weights [20]. Since, the dictionary $\mathbf{A}$ is trained on the live speech data, it is better suited to approximate the live speech utterances than their replayed counterparts. As a consequence, the error vectors from the sparse approximation provides an estimate of the device distortion component. Hence, the residual error vectors can be used as features for replay spoof detection. Given a feature vector $\mathbf{y}$ extracted from a test utterance, it can be represented using a sparse weight vector $\mathbf{x}$ and learned dictionary $\mathbf{A}$ by solving

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 , \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \tag{12}$$

using the orthogonal matching pursuit (OMP) algorithm [21]. The residual error vector in this approximation

$$\mathbf{e} = \mathbf{y} - \mathbf{A}\mathbf{x} \tag{13}$$

can be used as a feature to detect replayed speech.

The t-stochastic neighbourhood embedding (t-SNE) plots in Fig. 2 demonstrate the effectiveness of the proposed dictionary learning approach in extracting the device specific features. In the t-SNE plot of raw IFCC features (Fig. 2(a)), there is a significant overlap between the live and replayed speech utterances. However, in the t-SNE plot of residual errors (Fig. 2(b)), the replayed speech utterances are widely spread out, while the live speech utterances are concentrated around the origin. The concentration of live speech utterances around the origin can be attributed to the fact that the error vector $\mathbf{e}$ will be much lower for the live speech utterances. Hence, the density of the error vectors can be used to distinguish the replayed speech from live speech.

Given the live and replayed speech utterances, the residual error vectors are estimated using the overcomplete dictionary learned

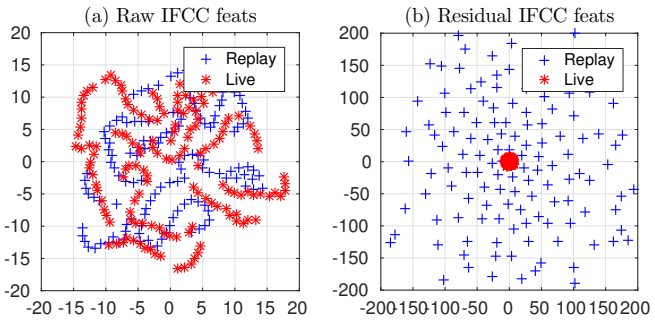

Fig. 2. Effect of highlighting device-specific characteristics. t-SNE plot of (a) raw IFCC features and (b) residual IFCC features.

from the live speech data. Two separate GMMs λ_L and λ_R are used to model the residual error vector distributions of live and replayed speech signals, respectively. Expectation maximization (EM) algorithm [22] is used to estimate the GMM parameters. Given a sequence of error vectors $\mathbf{E}$ extracted from an unknown test utterance, hypothesis testing is performed to classify it as either live or replayed speech. The confidence score is calculated as

$$\frac{P(\mathbf{E}/\lambda_R)}{P(\mathbf{E}/\lambda_L)} \underset{live}{\overset{replay}{\gtrless}} \eta \tag{14}$$

where $P(\mathbf{E}/\lambda_R)$ and $P(\mathbf{E}/\lambda_L)$ denote the likelihood data being generated from replayed and live speech models, respectively. The threshold η can be adjusted according to the required operating point. Too large a value of η leads to misdetections, while too small a value results in false alarms.

4. EXPERIMENTAL EVALUATION

The performance of the proposed IFCC based features was evaluated on the recently released ASVspoof 2017 corpus [23] as a part of the ASVspoof challenge 2017 [3]. The live speech data, consisting of RedDots utterances [24] of about 3-5 sec, was collected in diverse acoustic environments. Spoof data is created by replaying/rerecording these utterances on heterogeneous devices in different acoustic environments. All the speech data is sampled at 16 kHz. The dataset is divided into three sets - train, development (dev) and evaluation (eval) - each with disjoint set of speakers across them. The train set was used for building the models, and development set was used to fine-tune the hyper-parameters of the system. The performance of the overall system is evaluated on the blind evaluation dataset.

The speech signal is passed using a bank of 60 Gaussian shaped linearly-spaced filters with 200 Hz bandwidth to obtain multiple NB components. Smoothed IF is computed on each of these NB components, after taking a moving average of the numerator and denominator of (7) separately. The smoothed IF is averaged over frames of 25 ms, shifted by 10 ms, to obtain mean IF vector for each frame. IFCC features are extracted by applying DCT on mean IF vectors and retaining the first 20 coefficients in the DCT domain. The IFCC features are appended with their first and second order derivatives, that result in a 60-dimensional feature vector for further processing. In order to highlight the device-specific characteristics, an overcomplete dictionary with 1000 atoms was trained, on the IFCCs extracted from the live speech data, with a sparsity constraint parameter of $\tau = 5$, using K-SVD algorithm. The trained dictionary is used to estimate the live speech component from the IFCCs extracted from the replayed speech using OMP algorithm. The estimated live speech component is subtracted from replayed speech to obtain device-specific error vectors. The features extracted from live and replayed speech are modeled using two 512 mixture GMMs. Given a test utterance, hypothesis testing is performed to classify it as either live or replayed. The threshold parameter η in (14) is tuned using the development set. The sparsity constraint τ is optimized empirically using the development dataset. The effect of τ on the performance of spoof detection task is shown in Fig 3. Too small a τ is not sufficient to approximate the live speech component, whereas too large a value allows the channel distortion into the approximation. Hence, we have chosen $\tau = 5$ in all the further studies.

The performance of the proposed analytic phase features is compared with short-time spectral magnitude (MFCC & CQCC) and short-time spectral phase (MGDC) features. All the short-time spectral features are extracted from 25 ms of hamming windowed speech signal, shifted by 10 ms. In this study, we have used 13-MFCCs,

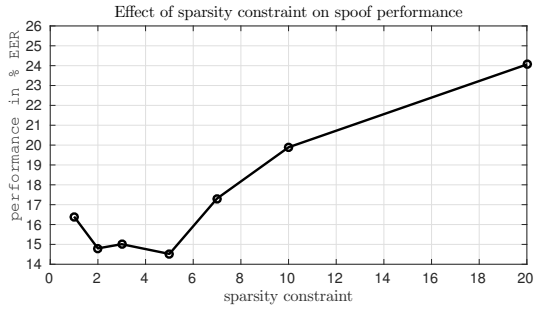

Fig. 3. Effect of sparsity on the performance of spoof detection

Table 1. Peformance of spoof detection system on ASVspoof - 2017 dataset, in terms of %EER

Feature	Raw features	Error features
CQCC	24.65	22.45
MFCC	30.48	21.4
MGDC	30.00	34.5
IFCC	**23.44**	**15.00**
MFCC+IFCC		**13.99**

30-CQCCs and 12-MGDCs along with their first and second order derivatives. For all the three features, we followed exactly the same modeling procedure as that of the IFCC features. The performance of the spoof detection system is evaluated in terms of equal error rate (EER), the point at which the false-acceptance and false-rejection rates are equal. The performance of different features on the evaluation set of the ASVspoof-2017 dataset is given in Table. 1. The CQCC/GMM sytem marked with $(*)$ is the baseline system supplied by the organizers of the ASVspoof challenge - 2017.

The proposed IFCC feature vectors outperform the other features, both in raw form as well as in error vector from. The error vectors extracted from the dictionary learning performed consistently better on all the feature types. In the case of MFCC and IFCC features, this improvement is substantial. i.e., in the order of 8% absolute improvement over the raw features. Hence, the evidence form the MFCC and IFCC error vector systems are combined to achieve the best performance of 13.99%, which is almost 10% relative improvement over the baseline system. These results substantiate that analytic phase captures the crucial information for discriminating live speech from the replayed speech.

5. CONCLUSION

This work highlights the importance of analytic phase in the context of replay attacks in ASV systems. The features derived from the instantaneous frequency, which is time derivative of phase, is used to capture the subtle acoustic variations in live and replayed speech. An overcomplete dictionary is learned from the features extracted from the live speech data. The residual components are obtained for a given utterance by subtracting the live speech contribution from the learned dictionary. GMMs are trained on the residual components from live and replayed speech. Hypothesis test based on a threshold provides the decision on whether a given utterance is live speech or playback from a recording device using the likelihoods produced from the respective GMMs. IFCC features perform better than magnitude based features such as MFCC and CQCCs and also other phase based features MGDCs based on group delay. The score fusion of MFCC and IFCC features outperforms the individual features as the magnitude and phase based features provide acoustic information from the complete spectrum. This clearly signifies the role of analytic phase in detection of replay attacks in automatic speaker verification systems.

6. ACKNOWLEDGEMENTS

The authors acknowledge the Ministry of Human Resource Development (MHRD) and the Ministry of Electronics and Information Technology (MeitY), Government of India, for sponsoring this work under IMPRINT project.

7. REFERENCES

[1] Yee Wah Lau, Michael Wagner, and Dat Tran, "Vulnerability of speaker verification to voice mimicking," in *Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on*. IEEE, 2004, pp. 145–148.

[2] Zhizheng Wu and Haizhou Li, "Voice conversion and spoofing attack on speaker verification systems," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*. IEEE, 2013, pp. 1–9.

[3] Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Kong Aik Lee, Md Sahidullah, Massimiliano Todisco, and Héctor Delgado, "Asvspoof 2017: automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training*, vol. 10, no. 1508, pp. 1508, 2017.

[4] 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.

[5] Jesús Villalba and Eduardo Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA workshop*, 2010, pp. 131–134.

[6] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *Speaker Odyssey Workshop, Bilbao, Spain*, 2016, vol. 25, pp. 249–252.

[7] Judith C Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[8] B. C. J. Moore, "An introduction to the psychology of hearing," *BRILL*, 2003.

[9] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," *Proc. Interspeech 2017*, pp. 2–6, 2017.

[10] Roberto Font, Juan M Espın, and Marıa José Cano, "Experimental analysis of features for replay attack detection–results on the asvspoof 2017 challenge," *Proc. Interspeech 2017*, pp. 7–11, 2017.

[11] Weicheng Cai, Danwei Cai, Wenbo Liu, Gang Li, and Ming Li, "Countermeasures for automatic speaker verification replay spoofing attack: On data augmentation, feature representation, classification and fusion," *Proc. Interspeech 2017*, pp. 17–21, 2017.

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

[13] B Shaik Mohammad Rafi, K Sri Rama Murty, and Shekhar Nayak, "A new approach for robust replay spoof detection in asv systems," in *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP) 2017*, pp. 51–55.

[14] Leon Cohen, *Time-frequency analysis*, vol. 778, Prentice hall, 1995.

[15] B Boshash, "Estimating and interpreting the instantaneous frequency of a signal," *Part*, vol. 2, pp. 540–568, 1992.

[16] Thomas F Quatieri, *Discrete-time speech signal processing: principles and practice*, Pearson Education India, 2006.

[17] Karthika Vijayan, Pappagari Raghavendra Reddy, and K Sri Rama Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication*, vol. 81, pp. 54–71, 2016.

[18] Leon Cohen, "Instantaneous frequency and group delay of a filtered signal," *Journal of the Franklin Institute*, vol. 337, no. 4, pp. 329–346, 2000.

[19] Vivek Boominathan and K Sri Rama Murty, "Speaker recognition via sparse representations using orthogonal matching pursuit," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4381–4384.

[20] Michal Aharon, Michael Elad, and Alfred Bruckstein, "rm K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[21] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, 1993*, pp. 40–44.

[22] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[23] Tomi Kinnunen, Md Sahidullah, Mauro Falcone, Luca Costantini, Rosa González Hautamäki, Dennis Thomsen, Achintya Sarkar, Zheng-Hua Tan, Héctor Delgado, Massimiliano Todisco, et al., "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[24] Kong Aik Lee, Anthony Larcher, Guangsen Wang, Patrick Kenny, Niko Brümmer, David van Leeuwen, Hagai Aronowitz, Marcel Kockmann, Carlos Vaquero, Bin Ma, et al., "The reddots data collection for speaker recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.