INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Energy Separation-based Instantaneous Frequency Estimation for Cochlear Cepstral Feature for Replay Spoof Detection

Ankur T. Patil[1], *Rajul Acharya*[1], *Pulikonda Aditya Sai*[2], *Hemant A. Patil*[1]

[1]Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India.
[2]Indian Institute of Information Technology (IIIT), Vadodara, Gujarat, India.

ankur_patil@daiict.ac.in, rajul_acharya@daiict.ac.in, 201551013@iiitvadodara.ac.in, hemant_patil@daiict.ac.in

Abstract

Replay attack poses significant threat to Automatic Speaker Verification (ASV) system among various spoofing attacks, as it is easily accessible by low cost and high quality recording and playback devices. This paper presents a novel feature set, i.e., Cochlear Filter Cepstral Coefficient Instantaneous Frequency using Energy Separation Algorithm (CFCCIF-ESA) to develop countermeasure against replay spoofing attacks. Experimental results on ASVspoof 2017 Version 2.0 database reveal that the proposed CFCCIF-ESA performs better than the earlier proposed CFCCIF (using analytic signal generation via Hilbert transform) feature set. This is because ESA uses extremely short window to estimate instantaneous frequency being able to adapt during speech transitions across phonemes. Experiments are performed using Gaussian Mixture Model (GMM) as a classifier. Baseline Constant Q Cepstral Coefficient (CQCC) performs slightly better than CFCCIF-ESA on development set (i.e., 12.47 % and 12.98 % Equal Error Rate (EER) for CQCC and CFCCIF-ESA, respectively). However, contrasting results on evaluation set (i.e., 18.81 % and 14.77 % EER for CQCC and CFCCIF-ESA, respectively) indicates that the proposed CFCCIF-ESA gives relatively better performance for unseen attacks in evaluation data. Also, the proposed feature set gives an EER of 11.56 % and 13.26 % on development and evaluation dataset when fused with state-of-the-art Mel Frequency Cepstral Coefficient (MFCC).

Index Terms: Auditory transform, ESA algorithm, cochlear filter, instantaneous frequency.

1. Introduction

Automatic Speaker Verification (ASV) system is used to verify claimed speaker's identity with the help of machines [1, 2]. Practically, an ASV system should be robust against variations, like, microphone and transmission channel, speaker aging, intersession, acoustic noise, etc. This will make the replayed speech more close to the natural speech as these variations are nullified. Subsequently, we want the framework to be secure against such spoofing attacks. Various kinds of spoofing attacks in voice biometrics are: voice conversion (VC) [3, 4], speech synthesis (SS) [5, 6], replay [7, 8], twins [9], and impersonation [10].

The ASVspoof campaign was initiated in 2015 to develop the countermeasures against spoofing attacks, which provides standard corpora, metrics, and protocol to support common evaluation. Three international challenges, namely, ASVspoof 2015, ASVspoof 2017, and ASVspoof 2019 have been organized so far to promote research in this direction [11, 12]. ASVspoof 2015 challenge was designed to develop countermeasures against SS and VC attacks, whereas ASVspoof 2017 challenge assessed various countermeasures for the replay attack. The ASVspoof 2019 challenge considers SS, VC, and replay attacks. Replay attack is easy to implement where the attacker imitates the genuine speaker by replaying the voice samples of target speaker. The replay speech recorded with a high quality recording and playback device in a clean recording environment is very hard to detect as it is very similar to genuine speech [13]. Hence, the present ASV systems are highly vulnerable to replay attacks. One of the way of the mathematical modelling of replayed speech is convolution of genuine speech signal with the impulse response of the recording environment and recording device, impulse response of the playback device, and playback environment [7, 14]. This modelling of replayed speech is used to develop countermeasures against replay attacks [15].

In [16, 17], auditory transform-based Cochlear Filter Cepstral Coefficient (CFCC) features were proposed to capture perceptual information in speech processing applications, such as classification of fricative sounds [18], and speaker identification [17]. However, the study in [19] shows that both phase and envelope of the cochlear filter are important for speech perception. Hence, in [19, 20], Cochlear Filter Cepstral Coefficient Instantaneous Frequency (CFCCIF) feature set was proposed, where both envelope and phase information were used for Spoofed Speech Detection (SSD) task. In [20], instantaneous frequency (IF) is estimated using the well known Hilbert transform (HT) method. However, HT is computationally expensive and has poor time resolution [21, 22]. To address this issue, in this paper we propose to use Teager Energy Operator (TEO) [23, 24] for IF estimation to exploit high time resolution property [25]. Thus, the resulting feature set, namely, Cochlear Filter Cepstral Coefficient Instantaneous Frequency Energy Separation Algorithm (CFCCIF-ESA) encapsulates phase information (using ESA) more effectively as compared to the original CFCCIF features. Results also suggest that CFCCIF-ESA performs better than the baseline Constant Q Cepstral Coefficient (CQCC) and CFCCIF on evaluation set of ASVspoof 2017 version 2 dataset [26].

2. Development of CFCCIF-ESA features

The CFCC feature extraction involves cochlear filter-based on auditory transform (AT), hair cell function, non-linearity, and discrete cosine transform (DCT) [17]. The CFCCIF features have both envelope and phase information. This might be the reason that CFCCIF performs better than the CFCC [20, 19]. The performance of the CFCCIF for replay SSD is further enhanced by estimating the instantaneous frequency (IF) by En-

http://dx.doi.org/10.21437/Interspeech.2019-2742

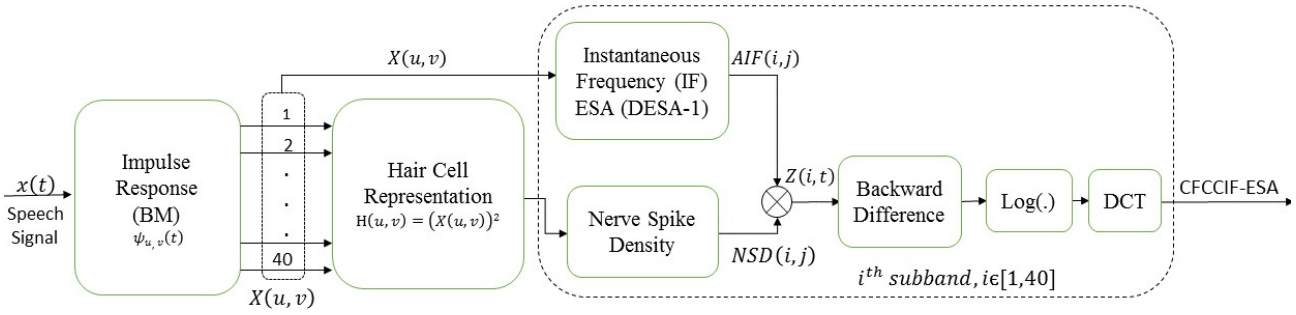

Figure 1: *Block diagram of the proposed CFCCIF-ESA. After [16, 17].*

ergy Separation Algorithm (ESA). The following sub-sections presents the computational technique of CFCC and proposed CFCCIF-ESA feature representation. An overview of the proposed CFCCIF-ESA feature extraction is depicted in Figure 1.

2.1. Auditory Transform (AT)

The AT models the frequency distribution of the cochlea in the human ear. It consists of the set of subband filters whose impulse responses describe the travelling wave in the cochlea. The AT ($X(u,v)$) of the signal $x(t)$ can be given as [17]:

$$X(u,v) = x(t) * \psi_{u,v}(t), \tag{1}$$

$$\psi_{u,v}(t) = \frac{1}{\sqrt{u}}\psi\Big(\frac{t-v}{u}\Big). \tag{2}$$

where $\psi(t)$ is the BM impulse response in the cochlea (also called as mother wavelet in signal processing literature.) [16, 17]. $*$ is convolution operation, $u \in R^+$, $v \in R$, $x(t)$ and $\psi(t) \in L^2(R)$. The factor u is scaling parameter which allows the modification in the central frequency, and v is called translation parameter. Here, $X(u,v)$ represents the travelling wave in the BM. The cochlear filter is defined as in [17]:

$$\begin{aligned}\psi_{u,v}(t) = \frac{1}{\sqrt{u}}\Big(\frac{t-v}{u}\Big)^{\alpha} \cdot exp\Big[-2\pi f_l \beta\Big(\frac{t-v}{u}\Big)\Big] \\ \times cos\Big[2\pi f_l\Big(\frac{t-v}{u}\Big) + \theta\Big]U(t-v),\end{aligned} \tag{3}$$

where $U(\cdot)$ is the unit step function. Parameters α and β regulate the shape and width of the frequency response of the cochlear filter, respectively. Parameter θ is selected such that $\psi(t)$ is function of zero average, i.e.:

$$\int_{-\infty}^{\infty} \psi(t)dt = 0 \Rightarrow \psi(\omega)|_{\omega=0} = 0. \tag{4}$$

It suggest that the $\psi(\omega)$ is bandpass in nature. It also ensures the existance of a number C_ψ such that $C_\psi = \int_0^\infty \frac{|\psi(\omega)|^2}{\omega} d\omega < \infty$, which satisfies the admissibility condition [27]. The value u varies for each filter in the filterbank and is calculated for each of the i^{th} subband filter. It is derived from the lowest frequency, f_l, and central frequency, f_c, of each filter of the cochlear filterbank [16]:

$$u = \frac{f_l}{f_c}. \tag{5}$$

2.2. Hair Cell Function

Subband filtering as explained above, emulates the impulse response in the cochlea. According to the frequency contents in the signal, BM regions move up and down. It causes the displacement of the hair cells to initiate the neural activity. However, inner hair cells generate the neural activity in single direction. To model the movement of inner hair cells in single direction, squared function can be used as:

$$H(u,v) = (X(u,v))^2, \quad \forall X(u,v), \tag{6}$$

where $X(u,v)$ is the filterbank output, and $H(u,v)$ is the hair cell function.

2.3. Nerve Spike Density

The hair cell output is then transformed to electrical signal, which is carried by auditory nerves to the brain [28]. The intensity of this electrical signal can be modeled by Nerve Spike Density (NSD) which is computed as follows:

$$NSD(i,j) = \frac{1}{l}\sum_{b=n}^{n+l-1} H(i,v), \quad n = 1, N, 2N, ...; \forall i, j, \tag{7}$$

where l is the window length, and N is the window shift duration. This output is further applied to the scales of loudness functions [29], which is logarithmic non-linearity in this case. After this, *discrete cosine transform* (DCT) is taken to get CFCC feature set [17].

2.4. Average Instantaneous Frequency (AIF) Estimation

The study reported in [20] shows that incorporating IF information with CFCC features enhances the representation. In earlier proposed CFCCIF feature set, IF is estimated using analytic signal generated via HT. The details of IF estimation and AIF computation using HT, are given in [19, 20]. In this paper, we are using ESA to estimate IF.

Teager's work on non-linear modelling of the human speech production in [23, 24, 30] was used to model, and detect the modulations in speech resonances [21, 31]. The ESA is based on a nonlinear differential operator, called as Teager Energy Operator (TEO), Ψ, which for the discrete-time signal $x(n)$ is given by Ψ_d [25, 32], i.e.:

$$\Psi_d\{x(n)\} = x^2(n) - x(n-1)x(n+1). \tag{8}$$

In [21], various *discrete energy separation algorithm* (DESA) are used for estimating the IF_i of i^{th} subband signal, which is given as:

$$IF_i = cos^{-1}\Big[1 - \frac{\Psi\{x_i(n) - x_i(n-1)\}}{2\Psi\{x_i(n)\}}\Big], \tag{9}$$

where $x_i(n)$ is the i^{th} subband signal. The above expression is a result of DESA-1. Next, framewise AIF for each of i^{th}

subband is obtained as:

$$AIF(i,j) = \frac{1}{l}\sum_{b=n}^{n+l-1} IF_i(b), \quad n = 1, N, 2N, ...; \forall i, j, \tag{10}$$

where l is the window length, and N is the window shift duration. In CFCCIF-ESA computation, ESA is used for IF estimation.

2.5. Estimation of CFCCIF-ESA Features

The nerve spike density gives the envelope structure while AIF gives the IF information of the windowed speech signal. Thus, product of the two encapsulates jointly both envelope and frequency information at each instant. Also, IF obtained in the silence regions is also suppressed. In particular, for each of the i^{th} subband using eq. (7) and eq. (10) , we obtain $Z(i,t)$ which is given by:

$$Z(i,t) = NSD(i,t) \times AIF(i,t). \tag{11}$$

To capture the transient information, change in resulting signal $Z(i,t)$ is captured using the derivative operation followed by logarithm. This operation is repeated for each of the i^{th} subband filter, $i \in [1, 40]$. Finally, DCT is applied framewise to get CFCCIF-ESA feature set.

3. Experimental Setup

3.1. Dataset

In this paper, ASVspoof 2017 Challenge Version 2.0 database is used which is a collection of natural (bonafide) and spoofed utterances. Bonafide utterances are a subset of the *RedDots* corpus [33]. These bonafide utterances are replayed and re-recorded to generate spoofed utterances. Recording of spoofed utterances was done using variety of heterogeneous devices, and acoustic environments. This corpus is divided into disjoint training, development, and evaluation sets. The details of the dataset is reported in [26].

3.2. Feature Set

In this paper, AT-based features, namely, CFCC, CFCCIF, and proposed CFCCIF-ESA are used along with baseline CQCC, and MFCC feature sets. We empirically found that, AT-based features (CFCC, CFCCIF, and proposed CFCCIF-ESA) are performing better for $\alpha = 3$, and $\beta = 0.019$ (eq. 3). These feature sets extracted using 40 linearly-spaced cochlear filterbanks. *12*-dimensional (*D*) static coefficients are extracted per utterance per filter excluding 0^{th} coefficient. These static features are appended with velocity (Δ) and acceleration ($\Delta\Delta$) coefficients to form *36-D* feature set. Baseline is implemented with *90-D* (static+Δ+$\Delta\Delta$) CQCC feature set. The state-of-the-art Mel Frequency Cepstral Coefficient (MFCC) feature set (*39*-D) is used for the comparison. Cepstral Mean and Variance Normalization (CMVN) is applied on the feature sets to enhance the performance of SSD system by eliminating the channel distortion [34]. A Gaussian Mixture Model (GMM) is used as a back-end classifier to distinguish speech utterances as natural or replayed. The decision of the final scores of speech signal being natural or replayed is decided by the Log-Likelihood Ratio (LLR), and is given by:

$$LLR = log\frac{P(X|N)}{P(X|R)}, \tag{12}$$

where $P(X|N)$, and $P(X|R)$ are the likelihood scores of natural and replay trials. The performance of the system is evaluated using Equal Error Rate (EER) which can be obtained using Detection Error Trade-off (DET) plot [35]. Score-level fusion of the two feature sets is performed as follows:

$$LLk_{fused} = \alpha \cdot LLk_{feature1} + (1-\alpha) \cdot LLk_{feature2}, \tag{13}$$

where $\alpha \in [0, 1]$. $LLk_{feature1}$ and $LLk_{feature2}$ are log-likelihood (LLk) scores of feature set-1 and feature set-2, respectively.

4. Experimental Results

4.1. Spectrographic Analysis

Figure 2 shows the spectrographic analysis of natural (in Panel I) *vs.* spoofed (in Panel II) speech signal for the same uttarance and speaker. Figure 2(b) and figure 2(c) display the spectral energy densities obtained from CFCCIF and CFCCIF-ESA feature sets, respectively. It can be clearly observed from the spectrograms of Panel I and Panel II that spoofed speech has relatively poor spectral energy density in the high frequency region (as seen by the dotted region in figure 2(b) and figure 2(c)). This is mainly due to the bandpass nature of frequency response of recording and playback devices and, frequency response of other environmental disturbances. It is also observed that enhanced spectral representation is obtained using CFCCIF-ESA as compared to that of CFCCIF feature set. This can be explained on the basis that DESA uses an extremely short window which allows it to make fast transitions across speech phonemes. HT on the other hand, needs a window whose length is of the same order as the length of the speech analysis frame [21]. Short window used in ESA algorithm to extract the IF, achieves better time resolution over HT-based approach. Hence, SSD system developed using CFCCIF-ESA shows percentage reduction in EER by 15.01% as that of CFCCIF counterpart.

4.2. Results on Individual Systems

In total 5 individual SSD systems were built using single feature set per classifier. The CMVN is applied on features in each SSD system to improve the performance by removing the channel distortion [36, 37, 38]. It seems unreasonable to eliminate the channel effect as channel information consists of the discriminative cues for the SSD task. However, given dataset consists of several configurations of the acoustic environment, recording and playback devices. Because of this wide range of

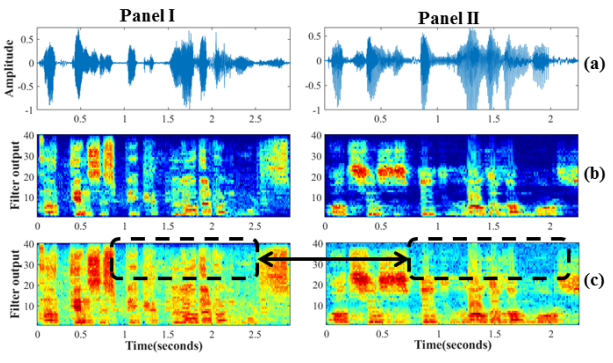

Figure 2: *Spectrographic analysis of genuine (Panel I) vs. spoofed (Panel II) speech signal : (a) time domain waveform and the corresponding spectral energy density (for 40 cochlear subband filters) of, (b) CFCCIF, and (c) CFCCIF-ESA.*

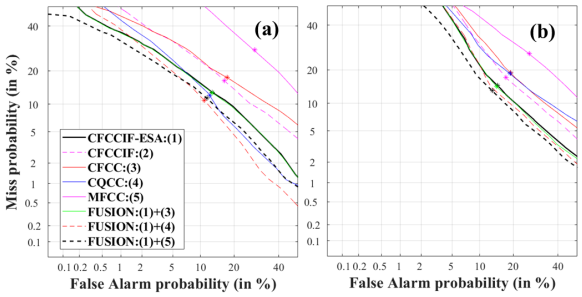

Figure 3: *DET curves for replay SSD systems. The individual DET curves for CFCCIF-ESA (proposed), CFCCIF, CFCC, CQCC, MFCC, and their score-level fusion on (a) development set, and (b) evaluation set. Legend shown in Fig. 3(a) is same for Fig. 3(b).*

channel variability, replayed speech signal might be difficult to classify. The CMVN normalizes this channel variability to get improved performance of the SSD system. The performance results of feature sets with GMM as a classifier is shown in Table 1. The baseline system is built using the *90-D* CQCC feature set. It shows 12.27 % and 18.81 % EER on development (Dev) and evaluation (Eval) set, respectively. The SSD system is also built up using the state-of-the-art MFCC feature set (*39-D*) which shows poor performance. The proposed CFCCIF-ESA feature set, shows comparable results on development set, whereas EER is reduced by 4.04 % (percentage reduction of 21.47 %) on evaluation set as compared to the baseline system. It shows the generalization capability of proposed feature set on unseen attacks. In addition, 2.61 % of absolute reduction (15.01 % of percentage reduction) is obtained relative to the CFCCIF feature set. This is because of the more accurate estimation of IFs using ESA.

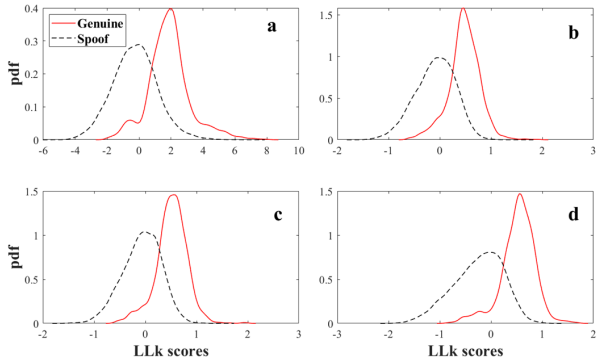

Figure 4: *The distribution of log-likelihood (LLk) scores on evaluation dataset for genuine samples (solid line) and spoof samples (dashed line) obtained from the SSD systems built using (a) CQCC, (b) CFCC, (c) CFCCIF, and (d) CFCCIF-ESA. Legend shown in Fig. 4(a) is same for Fig. 4(b),(c) and (d).*

4.3. Observation from DET Plots and Score Distributions

The DET curves of CFCCIF-ESA, CFCCIF, CFCC, CQCC, MFCC, and their possible fusion are shown in figure 3. It can be observed from figure 3(a) that SSD system implemented using CQCC is performing slightly better than proposed feature set on development data. Whereas, the DET curve in figure 3(b) shows that the performance of the SSD system developed using CFCCIF-ESA feature set is much better than its baseline CQCC and CFCCIF counterparts. In addition, significant decrease in miss probability is observed for CFCCIF-ESA on evaluation set, which is desirable property for the ASV system. The quantity is further reduced when score-level fusion of proposed system CFCCIF-ESA is performed with the baseline CQCC and MFCC systems. Figure 4 shows log-likelihood score distribution on evaluation set for genuine (solid red line), and spoof (dashed black line) samples for SSD systems using the feature sets. Figure 4(a) shows the score distribution of baseline CQCC system, whereas Figure 4(b), (c) and (d) shows the score distribution of CFCC, CFCCIF, and CFCCIF-ESA systems, respectively. It is clearly observed that the area of overlapping region of score distribution of genuine and spoof samples is relatively lesser for CFCCIF-ESA system than that of other SSD systems.

Table 1: *Results on development (Dev) and evaluation (Eval) dataset for individual systems trained on GMM*

Feature Sets	# Gaussian Mixtures	% EER	
		Dev	Eval
CQCC (baseline)	512	12.27	18.81
MFCC	512	28.52	26.45
CFCC	512	17.60	18.97
CFCCIF	512	16.61	17.38
CFCCIF-ESA *	256	12.98	**14.77**
CFCCIF-ESA	512	12.92	14.87
CFCCIF-ESA* $\oplus$ CQCC	-	10.91	13.64
CFCCIF-ESA* $\oplus$ MFCC	-	11.56	**13.26**
CFCCIF-ESA* $\oplus$ CFCC	-	12.92	14.45
CFCCIF-ESA* $\oplus$ CFCCIF	-	12.67	14.24

where $\oplus$ denotes the score level fusion and * denotes the CFCCIF-ESA system trained using 256 Gaussian mixtures

4.4. Results on Score-Level Fusion of Feature Sets

To explore the possible complementary information in feature sets against proposed feature set CFCCIF-ESA, we have performed score-level fusion of CFCCIF-ESA system with the systems using other feature sets. The results reported in Table 1 shows that least EER of 13.26 % is obtained when score-level fusion of proposed feature set is performed with MFCC. MFCC captures the psychological aspects of speech perception, whereas CFCCIF-ESA captures the physiological aspects. This might be the reason that, their score-level fusion is showing reduced EER by capturing possible complementary information. It can be observed that the DET curves are more inclined downward for this fusion. This is desirable property for ASV system as the risk of spoofed sample recognized as genuine sample is reduced.

5. Summary and Conclusions

This paper presented the usefulness of the auditory transform-based features (CFCC) in conjunction with instantaneous frequency (IF) estimated using ESA algorithm (CFCCIF-ESA) for replay SSD system development. The performance of the proposed CFCCIF-ESA feature set is compared with the CQCC (baseline), MFCC, CFCC, and CFCCIF. The performance of the proposed feature set is eloquent over other feature sets. The earlier proposed CFCCIF feature set has used Hilbert transform-based approach to estimate the IF. Whereas, proposed feature set uses TEO-based ESA algorithm to estimate IF which significantly improves the spectrographic representation of the speech signal and hence, the proposed feature set perfoms better than the CFCCIF feature set for replay SSD task.

6. References

[1] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[2] S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans, *Handbook of Biometric Anti-spoofing*, 2^{nd} ed. Springer, 2018.

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[4] Y. Stylianou, "Voice transformation: A survey," in *ICASSP*, Taipei, Taiwan, April 2009, pp. 3585–3588.

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

[6] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2280–2290, 2012.

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

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

[9] 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.

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

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

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

[13] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Seim Reap, Combodia, December 2014, pp. 1–5.

[14] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *The Journal of the Acoustical Society of America (JASA)*, vol. 45, no. 2, pp. 458–465, 1969.

[15] P. A. Tapkir, A. T. Patil, N. Shah, and H. A. Patil, "Novel spectral root cepstral features for replay spoof detection," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Honolulu, Hawaii, USA, 2018, pp. 1945–1950.

[16] Q. Li, "An auditory-based transfrom for audio signal processing," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 181–184.

[17] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1791–1801, 2011.

[18] N. Singh, N. Bhendawade, and H. A. Patil, "Novel cochlear filter based cepstral coefficients for classification of unvoiced fricatives," *Int. J. Natural Lang. Comput*, vol. 3, no. 4, pp. 21–40, 2014.

[19] T. B. Patel and H. A. Patil, "Cochlear filter and instantaneous frequency based features for spoofed speech detection," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, 2017, pp. 618–631.

[20] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural *vs*. spoofed speech," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2062–2066.

[21] P. Maragos, J. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[22] A. Potamianos and P. Maragos, "A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation," *Signal Processing*, vol. 37, no. 1, pp. 95–120, 1994.

[23] H. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 599–601, 1980.

[24] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," *Speech Production and Speech Modelling, Springer*, pp. 241–261, 1990.

[25] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Mexico, USA, 1990, pp. 381–384.

[26] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. Lee, and J. Yamagishi, "ASVspoof 2017 version 2.0: metadata analysis and baseline enhancements," in *Odyssey 2018: The Speaker and Language Recognition Workshop*, 2018.

[27] S. Mallat, *A Wavelet Tour of Signal Processing*, 2^{nd} ed. Elsevier, 1999.

[28] B. C. Moore, *An Introduction to the Psychology of Hearing*. Brill, 2012.

[29] S. S. Stevens, "On the psychophysical law." *Psychological Review*, vol. 64, no. 3, p. 153, 1957.

[30] P. Maragos, J.F. Kaiser and T.H. Quatieri , "On separating amplitude from frequency modulations using energy operators," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Francisco, California, USA, 1992, pp. 1–4.

[31] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

[32] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. 1^{st} edition, Pearson Education India, 2015.

[33] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. v. Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma *et al.*, "The RedDots data collection for speaker recognition," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2996–3000.

[34] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.

[35] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *EUROSPEECH*, Rhodes, Greece, 1997, pp. 1895–1898.

[36] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection-results on the ASVspoof 2017 challenge." in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 7–11.

[37] 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 ASVspoof2017," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 87–91.

[38] 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.