Cochlear Filter and Instantaneous Frequency Based Features for Spoofed Speech Detection

Tanvina B. Patel, Student *Member, IEEE*, and Hemant A. Patil, *Member, IEEE*

***Abstract*—Vulnerability of voice biometrics systems to spoofing attacks by Synthetic Speech (SS) and Voice Converted (VC) speech have arose the need of standalone Spoofed Speech Detection (SSD) systems. The present work is an extension of our previously proposed features (used in relatively best performing SSD system) at the first ASVspoof 2015 challenge held at INTERSPEECH 2015. For the challenge, the authors proposed novel features based on Cochlear Filter Cepstral Coefficients (CFCC) and Instantaneous Frequency (IF), i.e., CFCCIF. The basic motivation behind this is that human ear processes speech in subbands. The envelope of each subband and its IF is important for perception of speech. In addition, the transient information also adds to the perceptual information that is captured. We observed that subband energy variations across CFCCIF when estimated by symmetric difference (CFCCIFS) gave better discriminative properties than CFCCIF. The features are extracted at frame-level and Gaussian Mixture Model (GMM)-based classification system was used. Experiments were conducted on ASVspoof 2015 challenge database with MFCC, CFCC, CFCCIF and CFCCIFS features. On the evaluation dataset, after score-level fusion with MFCC, the CFCCIFS features gave an overall Equal Error Rate (EER) of *1.45 %* as compared to *1.87 %* and *1.61 %* with CFCCIF and CFCC, respectively. In addition to detecting the known and unknown attacks, intensive experiments have been conducted to study the effectiveness of the features under the condition that either only SS or only VC speech is available for training. It was observed that when only VC speech is used in training, both VC, as well as SS, can be detected. However, when only SS is used in training, VC speech was not detected. In general, amongst vocoder-based spoofs, it was observed that VC speech is relatively difficult to detect than SS by the SSD system. However, vocoder-independent SS was toughest with highest EER (i.e., > *10 %*).**

***Index Terms*—Anti-spoofing, CFCC, instantaneous frequency, CFCCIFS, Gaussian mixture model.**

I. Introduction

Voice as a biometric modality has received a lot of attention over the years [1], [2]. The Automatic Speaker Verification (ASV) system is a voice biometric technology that uses speaker-specific properties from the speech signal for authentication purpose (to accept or reject a claimed speaker's identity) [2]. Although current ASV systems offer high accuracy and low % Equal Error Rate (EER), the ASV systems need to be reliable under spoofing scenarios as well. Spoofing can be due to impersonation (mimicking or voice disguise in speaker forensics), replay, speech synthesis and voice conversion. The ASV systems are known to be highly vulnerable to replay attacks. However, the ASV systems are also found to be susceptible to Text-to-Speech (TTS) synthesis systems (generally, Hidden Markov Model (HMM)-based TTS systems (HTS) [3]- [4] and adapted HMM-based systems [5]) and voice conversion attacks [6]- [7]. This is because of the advancements made in the development of TTS and voice conversion techniques that make them sound more natural and intelligible. In addition, the TTS and voice conversion systems can be easily developed by available open source techniques. A detailed description of previous studies on the effect of various spoofing attacks is presented in [8].

Initial work to identify the vulnerability of ASV systems to HMM-based Synthetic Speech (SS) was reported in [9], where the False Acceptance Rates (FAR) for ASV systems when spoofed by SS, reached to *70 %*. Even after using excitation source features in the ASV system, the FAR was around *20 %* [10]. Several studies show that with the use of adapted speech synthesis techniques for spoofing, the FAR of ASV systems can increase from *0-1 %* to *80-90 %* [11]- [12]. The Voice Converted (VC) speech when used as a spoof is also known to have similar or rather worse effect on the ASV systems. The VC spoof is shown to significantly increase the % EER for simple Gaussian Mixture Model (GMM)-based ASV systems [13]- [14]. For Joint Factor Analysis (JFA)-based ASV systems and *i*-vector-based ASV systems [15], the FAR has increased more than *5*-folds [16].

Earlier approach to detect natural and SS spoof was based on using instability in pitch patterns [17]. Later in [18], temporal modulation features were used to detect SS spoof. A well-known work in this area is based on using Relative Phase Shift (RPS) that demonstrates reliable detection of SS to improve the security of ASV systems [19]. The study reported in [20], continued comprehensive evaluation using RPS and state-of-the-art Mel Frequency Cepstral Coefficients (MFCC) to develop a standalone SS detector. For detection of VC speech, Modified Group Delay (MGD)-based phase features were used in [21]. These features were then used for a spoof detection system in the ASV framework [22]. In [23], VC

This work was supported in part by the Department of Electronics and Information Technology (DeitY), Government of India, New Delhi, through two consortium projects, TTS Phase-II and ASR Phase-II, and in part by the authorities of DA-IICT, Gandhinagar, India.

Tanvina B. Patel and Hemant A. Patil are with the Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar 382007, Gujarat, India (e-mail: tanvina_bhupendrabhai_patel@daiict.ac.in; hemant_patil@daiict.ac.in).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2016.2647201, IEEE Journal of Selected Topics in Signal Processing

anti-spoofing was performed using back-end models jointly with SV in the *i*-vector space. Most of the previous countermeasures used different spoofing databases and the discriminative features were applied to known attacks (i.e., using known prior information about the spoofing algorithm).

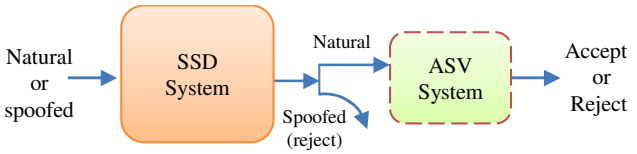


Fig. 1. A Spoofed Speech Detection (SSD) system for ASV system.

Previously, limited work considered the realistic scenario of unknown attacks or mismatched conditions [23]. Hence, it becomes difficult to determine the impact of a particular spoof on the ASV system and the efficiency of a particular anti-spoofing technique. To that effect, the Spoofing Anti-Spoofing (SAS) database [24] and the first ASVspoof 2015 challenge has been organized as a special session at INTERSPEECH 2015 [25]. For the ASVspoof 2015 challenge, the task was to design a standalone Spoofed Speech Detection (SSD) system that could classify natural and spoofed speech for both *known* and *unknown* attacks (as shown in Fig. 1). A spoofing dataset was provided and the results in % EER on the evaluation data were returned by the organizers. The training and development set consisted of vocoder-dependent spoofed speech while the evaluation set included both vocoder-dependent and vocoder-independent speech. The effect of the spoofing attacks on the % EER and % FAR of ASV systems is reported in [24], [26]. Various countermeasures are proposed till date and evaluated on the challenge data. Majority of these were phase-based, including MGD-based features [27], [28], relative phase features [29], [30], [31], etc. In addition, Linear Prediction (LP) residual-based features [32], [33], [34], wavelet-based features [35], subband processing based Linear Frequency Cepstral Coefficients (LFCC) [36]- [37] and the Constant Q Cepstral Coefficients (CQCC) feature sets [38], and other source-based features [39] were also introduced. Moreover, the *i*-vector-based systems [40], Deep Neural Networks (DNN)-based representation [41], [42] and the use of DNN and Support Vector Machine (SVM) classifier [43] for spoofing detection were explored.

For the ASV spoof challenge, the authors proposed an SSD system based on envelope and phase features. The basic idea is that the human ear processes speech in subbands primarily due to place-specific movement of the Basilar Membrane (BM). The envelope of the output of each cochlear filter and its phase are important features used by auditory levels for speech perception (Chapter 8, pp. 403 [44]). To capture the envelope-based features, we use the Cochlear Filter Cepstral Coefficients (CFCC) derived from the wavelet transform-like auditory transform [45]- [46]. The use of auditory-based filters in place of triangular filters in MFCC aids CFCC features to capture perceptual information. In addition, to capture the phase representation, the subband Instantaneous Frequency (IF) is estimated. The average IF representation is embedded into the CFCC framework to give CFCC plus IF (i.e., CFCCIF) features for the SSD task [47].

The organization of the rest of the paper is as follows. Section II describes the basis of the proposed work. Section III describes details of CFCC, CFCCIF and proposed CFCCIFS feature sets for SSD task. Section IV introduces the SSD architecture, the database and evaluation metrics used. Section V and Section VI shows the experimental results on the development and evaluation sets, respectively. Finally, Section VII concludes the paper along with future research directions.

II. Basis of the Proposed Approach

The usefulness of the auditory-based CFCC features in capturing perceptual information has been explored in speech processing applications such as speaker identification [46] and classification of fricatives sounds [48]. Earlier in [49], an auditory-based distortion measure was used to find the perceptual dissimilarity between speech segments and improve the quality of synthesized speech by selecting speech sound units based on the auditory distortion measures. Next, considering few applications of IF, in a study reported in [50] the short-time IF spectrum was found to contribute to the speech intelligibility as much as the short-time magnitude spectrum. In [51], the subband IF is used with the envelope from subband filter outputs for speech recognition task. Thus, features derived from phase may be supplementary to that of the features derived from magnitude spectrum. On the similar lines, for the application of SSD, we propose to jointly use the cochlear filter envelope-based and IF features.

In addition to using envelope and phase features, for the present problem of SSD, the authors are motivated to use the variations across frames of the CFCCIF representation [47]. Prior work shows that, countermeasures are designed based on the observation of different dynamic variation in the speech parameters of SS and natural speech. In [52], use of Intraframe Differences (IFD) as a discriminative feature was used due to the fact that in the HMM-based speech synthesis, the speech parameter sequence is generated to maximize the output probability and hence, the variation in likelihood will be less as compared to natural speech. In [53], higher-order Mel Cepstral Coefficients (MCEP) of SS revealed less variance than that of natural speech signal. This is because, the higher-order MCEP are smoothed during HMM model parameter training and synthesis. Next, as the feature extraction process in SS and VC speech generation is framewise, we use derivative operation to capture transient variations across the frames to assist in the SSD task. The use of derivative enhanced the high frequency regions in the representations which is a possible reason for improved performance. In [36], [54] similar observations were made that the higher frequency regions of speech are essential for the SSD task. In [47], a one-point backward difference was used, however, in this paper, the authors use *symmetric* difference to extract feature variations information across frames. We refer to this feature set as CFCCIFS. The symmetric difference (which uses both past and present sample) smoothens out abruptness due to one

sample differentiation (as in backward difference). It also intuitively represents the fact that both past and future context information are essential to perceiving transient information at a particular time instant, resulting in better SSD performance.

The CFCCIF features were found to be the relatively best performing features at the ASVspoof 2015 challenge and the SSD system proposed from DA-IICT team gave the least % EER over known and unknown attacks. This implicitly compares the CFCCIF features with other earlier techniques existing in the literature. The present work extends our preliminary work on CFCCIF [47]. In this work, as in several other studies [20], [31] we use MFCC features as baseline. The relative performance of MFCC, CFCC, CFCCIF and CFCCIFS feature set is compared. A GMM-based classifier is used as the back-end and the performance is evaluated on the basis of EER and Detection Error Trade-off (DET) curve. As compared to [47], in this paper, we improve upon the existing CFCCIF features and perform intensive evaluation to validate the efficiency and robustness of the features. The initial experiments are carried out to decide the number of subband filters and the effect of pre-emphasis in using MFCC and cochlear-based features. Effect of various window lengths in estimating the dynamic features from the static features is studied along with the individual contribution of static and dynamic features in the SSD task. In contrast to reporting the performance in the conventional sense of known attacks and unknown attacks, we evaluate the countermeasures by its ability to detect its '*same type*' of spoof and '*different type*' of spoof. That is, we train the GMM model on one technique (e.g., SS) and evaluate its performance to detect SS (i.e., same type) and VC (i.e., different type). This presents dependency of the countermeasures to the spoofing type (SS or VC). It is validated through experiments that for vocoder-based spoofs, the system trained on VC can detect both VC and SS. However, system trained on SS can detect only SS and not VC spoof. To the best of authors' knowledge, this is the first study to report such observations. In this setup, amongst all the feature sets, CFCCIFS features detected the same type and different type of spoofed speeches much better than rest of the features. Even in terms of the overall EER, the CFCCIFS feature set is found to work well amongst all the features.

III. Proposed Cochlear Filter and IF-based Features

This Section describes in brief the auditory transform and the method for estimating the CFCC features followed by IF estimation and procedure to use CFCC and IF features to obtain CFCCIF and the proposed CFCCIFS feature set for the SSD task.

A. Cochlear Filter Cepstral Coefficients (CFCC)

The parameter extraction procedure for auditory-based cepstral coefficients, consists of series of cochlear filterbank based on the auditory transform, hair cell function, nonlinearity and Discrete Cosine Transform (DCT) [46]. Following sub-Section defines in brief the auditory transform and the procedure to estimate CFCC features.

1) Auditory Transform

The auditory transform has well defined wavelet properties with an existing inverse transform [45]. It converts the time-domain signal into a set of filterbank output with frequency responses similar to those in the BM of the cochlea. Let $s(t)$ be the speech signal and the cochlear filter be $\psi(t)$. Thus, the auditory transform of $s(t)$ (i.e., $W(a,b)$), with respect to $\psi(t)$ as the impulse response of BM in the cochlea is defined in [45]- [46] as follows:

$$W(a,b) = s(t) * \psi_{a,b}(t), \tag{1}$$

$$W(a,b) = \int_{-\infty}^{\infty} s(\tau)\psi_{a,b}^{*}(t-\tau)d\tau, \tag{2}$$

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right). \tag{3}$$

where in eq. (1), * indicates convolution operation, $a \in \mathrm{R}^{+}$ and $b \in \mathrm{R}$, $s(t)$ and $\psi(t)$ belongs to Hilbert space $L^2(\mathrm{R})$ and $W(a,b)$ represents traveling waves in the BM. The factor a is the scale or dilation parameter, which allows changing the center frequency, f_c, while factor b is the time shift or translation parameter. The energy remains equal for all a and b. Hence, we have,

$$\int_{-\infty}^{\infty} |\psi_{a,b}(t)|^2 \, dt = \int_{-\infty}^{\infty} |\psi(t)|^2 dt. \tag{4}$$

The cochlear filter is defined as [46],

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\left(\frac{t-b}{a}\right)^{\alpha} \exp\left[-2\pi f_L \beta\left(\frac{t-b}{a}\right)\right] \times \cos\left[2\pi f_L\left(\frac{t-b}{a}\right) + \theta\right] u(t-b). \tag{5}$$

Parameters α and β determine the *shape* and *width* of cochlear filter and θ is selected such that the following *admissibility* condition for mother wavelet (i.e., $\psi(t)$) is satisfied [55]:

$$\int_{-\infty}^{\infty} \psi(t)dt = 0 \Rightarrow \psi(\omega)\big|_{\omega=0} = 0. \tag{6}$$

Thus, $\exists$ a number C_ψ such that, $C_\psi = \int_0^{\infty} \frac{|\psi(\omega)|^2}{\omega} d\omega < \infty$ (Theorem 4.3, pp. 81, [55]). This means that the mother wavelet $\psi(t)$ is a *bandpass* filter. The value of a can be derived from the central frequency, f_c, and the lowest frequency, f_L, of the cochlear filterbank, i.e.,

$$a = \frac{f_L}{f_c}. \tag{7}$$

For the i^{th} subband filter, its corresponding value of a, i.e., $\{a_i\}$ is pre-calculated for the required central frequency of the cochlear subband filters at band number $i \in [1, N_F]$, where N_F is the total number of subband filters.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2016.2647201, IEEE Journal of Selected Topics in Signal Processing

2) Other operations in CFCC extraction

Once filtering process is done by the cochlea in the ear, the inner hair cell acts as a transducer for the movements of BM. As motion of the hair cell is only in the *positive* direction, the following nonlinear function of hair cell describes this motion [46], i.e.,

$$h(a,b) = (W(a,b))^2; \qquad \forall W(a,b) \ , \tag{8}$$

where $W(a,b)$ is the filterbank output. The hair cell output of each filterbank is converted into a representation of the nerve spike density, which is computed as average of $h(a,b)$,

$$S(i,j) = \frac{1}{d}\sum_{b=l}^{l+d-1} h(i,b), \qquad l = 1, L, 2L, ...; \ \forall i, j \ , \tag{9}$$

where d is the window length, i is the i^{th} subband, j is the frame count and L is the window shift duration. The output of eq. (9), i.e., $S(i, j)$ is further applied for scales of loudness functions such as the *logarithmic* or *cubic root* nonlinearity. As per CFCC feature extraction in [46], we use the logarithmic nonlinearity. Finally, DCT is applied to decorrelate the features and get the CFCC feature vector.

B. Average Instantaneous Frequency (AIF) Estimation

In CFCC, the nerve spike density performs averaging operation on each subband signal which in turn removes the Temporal Fine Structure (TFS) or fast temporal modulations as in Fig. 2 (c) [56]. Furthermore, at every Characteristic Frequency (CF) of the cochlear filter (i.e., center frequency of the cochlear filter, f_c, as in eq. (7)), rapid phase shift of the travelling wave occurs at every f_c from base to apex of the BM [57]. We believe that this rapid change is being captured by derivative of instantaneous (analytic) phase (which is referred to as IF) of corresponding subband signal. For the SSD task, in vocoder-dependent spoofs, the phase information is generally lost. On the other hand, for vocoder-independent spoofs the phase mismatch and temporal discontinuity occurs due to joining of units. Hence, we propose to use the average IF for every subband along with the envelope representation obtained in the CFCC framework.

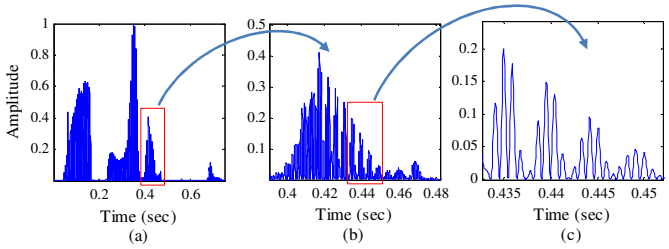


Fig. 2. For a subband around f_c=*550* Hz, (a) the slow modulations roughly correlate with the different segments of the utterance, (b) modulations due to interharmonic interactions occur at a rate that reflects the fundamental of the signal and (c) fast temporal modulations are due to the frequency component driving this subband around *550* Hz. After [56].

Hence, for the IF estimation, let $x_i(t)$ be the signal for the i^{th} subband. For the real signal $x_i(t)$, its complex *analytic* representation is given by,

$$x_{a_i}(t) = x_i(t) + jx_{h_i}(t), \tag{10}$$

where $x_{hi}(t)$ is the Hilbert transform of the signal $x_i(t)$, given by the inverse Fourier transform of $X_{hi}(\omega)$, where,

$$X_{h_i}(\omega) = \begin{cases} +jX_i(\omega) & \omega < 0 \\ -jX_i(\omega) & \omega > 0 \end{cases}. \tag{11}$$

Thus, the amplitude (Hilbert) envelope of $x_i(t)$ and the instantaneous phase for the i^{th} subband is given as,

$$| x_{a_i}(t) | = \sqrt{x_i^2(t) + x_{h_i}^2(t)} \text{ and } \phi_i(t) = \tan^{-1}\left(x_{h_i}(t) \Big/ x_i(t) \right). \tag{12}$$

Therefore, for the i^{th} subband, the IF derived from derivative of *unwrapped* instantaneous phase $\phi_i(t)$, is given as:

$$IF_i = \frac{d}{dt}(\phi_i(t)). \tag{13}$$

Next, similar to nerve spike density estimation, the framewise average IF for each i^{th} subband is obtained as,

$$AIF(i,j) = \frac{1}{d}\sum_{b=l}^{l+d-1} IF_i(b), \qquad l = 1, L, 2L, ...; \ \forall i, j \tag{14}$$

where d is the window length, j is the frame count and L is the window shift duration.

C. The CFCCIF and CFCCIFS Features

For each subband, the envelope (estimated in eq. (9)) needs to be combined with the average IF (estimate in eq. (14)). In particular, for each of the i^{th} subband, using eq. (9) and eq. (14), we have,

$$z(i,t) = S(i,t) \times AIF(i,t), \tag{15}$$

where $z(i,t)$ is the representation obtained after multiplying subband envelope and average IF features for the i^{th} subband. In [51], the subband IF was used explicitly by concatenating it with the envelope from subband filter outputs for the task of speech recognition. However, with this the feature dimension increases to twice. In [58], multiplication of envelope and fine structures estimated from the Hilbert transform of bandpass filtered signal was carried out. This was done to investigate the relative perceptual importance by chimaera synthesis. Therefore, we use multiplication of both envelope and average IF features to preserve the relevant information at the same feature dimension. Furthermore, the multiplication operation will suppress the random IF estimated in silence regions by the low amplitude values of the envelope structure.

Next, the derivative operation on $z(i,t)$ is performed (as shown by dashed block in Fig. 3) which can be represented as follows:

$$\therefore \frac{\partial(z(i,t))}{\partial t} = AIF(i,t)\frac{\partial S(i,t)}{\partial t} + S(i,t)\frac{\partial AIF(i,t)}{\partial t}, \tag{16}$$

Thus, the derivative of $z(i,t)$ representation is the sum of changes in nerve spike density weighted by average IF and the changes in average IF weighted by the nerve spike density. These features were proposed in [47], and are referred to as CFCCIF representation that uses backward difference for the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2016.2647201, IEEE Journal of Selected Topics in Signal Processing

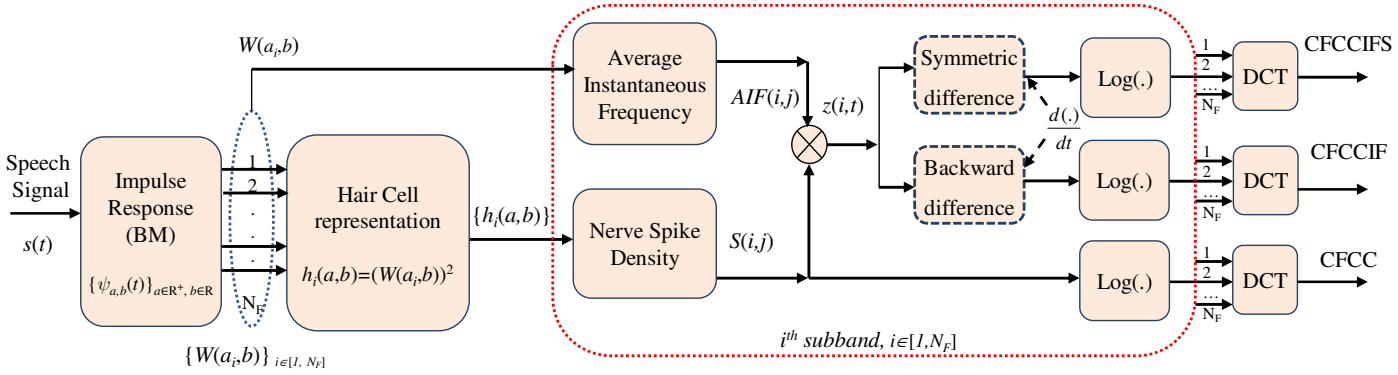


Fig. 3. Block diagram for CFCC, CFCCIF and proposed CFCCIFS feature extraction scheme.

derivative computation. In the present work, to capture both past and future context information at a particular time instant, we use the *symmetric* difference to estimate the change in CFCCIF features across the frames. The symmetric difference is used as follows,

$$\therefore \frac{\partial(z(i,t))}{\partial t} \approx \frac{z[i,n+1]-z[i,n-1]}{2}, \tag{17}$$

where $z(i,t)$ is given as per eq. (15). This is followed by logarithm and DCT is applied framewise to *decorrelate* the features for all the subbands, i.e., $i \in [1,N_F]$, (shown by dotted block in Fig. 3). The modified CFCCIF obtained by using the symmetric difference is denoted as the proposed CFCCIFS feature set.

D. *Effectiveness of Derivative Operation*

As shown in the previous sub-Section, instead of directly using the representation obtained by multiplying the subband envelope and IF, the change across the frames computed via derivative operation is used. The efficiency of the proposed features lies in exploiting the *dynamic* information via derivative operation for the SSD task. The effectiveness of the derivative operation is illustrated in Fig. 4. In particular, Panel I, Panel II, Panel III, and Panel IV considers the analysis of natural speech, vocoder-dependent SS, vocoder-dependent VC and vocoder-independent Unit Selection Synthesis (USS)-based speech, respectively, corresponding to the same text material from the SAS database [24]. Fig. 4(a) shows the speech waveform, Fig. 4(b) and Fig. 4(c) shows the subband energy output during the computation of MFCC and CFCC, respectively. Fig. 4(d) shows the subband energy output representation obtained on multiplying envelope and average IF (without derivative operation). Fig. 4(e) shows the subband energy representation of CFCCIF using backward difference operation as in [47] and Fig. 4(f) shows subband energy representation of CFCCIFS where the symmetric difference is used for derivative approximation.

Here, we study the effect of the envelope and the average IF features using a narrow *-3* dB bandwidth (high quality factor) cochlear filter with $\alpha = 3$ and $\beta = 0.035$. Narrower filters are needed for efficient IF estimation. Moreover, in the early auditory processing model of Shamma [56], [57], high quality cochlear subband filter responds only to frequencies near the center frequencies and hence, are found to produce more regular (i.e., periodic) synchronized responses even independent of input stimuli (such as noise, harmonic sequence or impulse) [57].

The spectrum of the auditory transform is known to preserve the formant information with less pitch harmonics (i.e., fundamental frequency, F_0) and computational noise [46]. It is observed in Fig. 4 (c) that the formant characteristics are enhanced more in natural speech than the vocoder-based SS and VC speech. The higher formants are an attribute of the natural speech and it is difficult to incorporate it in the machine-generated speech. In case of USS-based speech, the formant information is intact due to concatenation of natural speech sound units. However, this depends on the sound units that are picked. In Fig. 4(d), the representation obtained by multiplication of envelope and average IF features (without derivative) is similar to Fig. 4(c) except that the frequency regions are enhanced due to the embedded IF information. For spoofed speech in Panel II and Panel III, the high frequency regions are enhanced after multiplication with average IF. Next, the CFCCIF filterbank representation as in [47] is shown in Fig. 4(e) for all the panels. After taking derivative, the features corresponding to natural and spoofed speech have been more discriminative. For natural speech, the variations of envelope and average IF across the frames were more visible along all the filterbanks as compared to that of the SS and VC speech. Furthermore, by using symmetric difference, the subband energy representation is smoother and in fact, the difference between natural and spoofed speech (i.e., SS and VC) is much more prominent. Therefore, from embedding the average IF information to taking the derivative, the high frequency regions have significantly enhanced as shown by dotted regions in Fig. 4. In recent studies, it has been observed that high-frequency regions indeed are essential for spoof detection [36], [54]. Along the similar lines, the performance of CFCCIF and CFCCIFS is found to be better than the CFCC features which will be further quantified through intensive experiments presented in Section V and Section VI.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2016.2647201, IEEE Journal of Selected Topics in Signal Processing

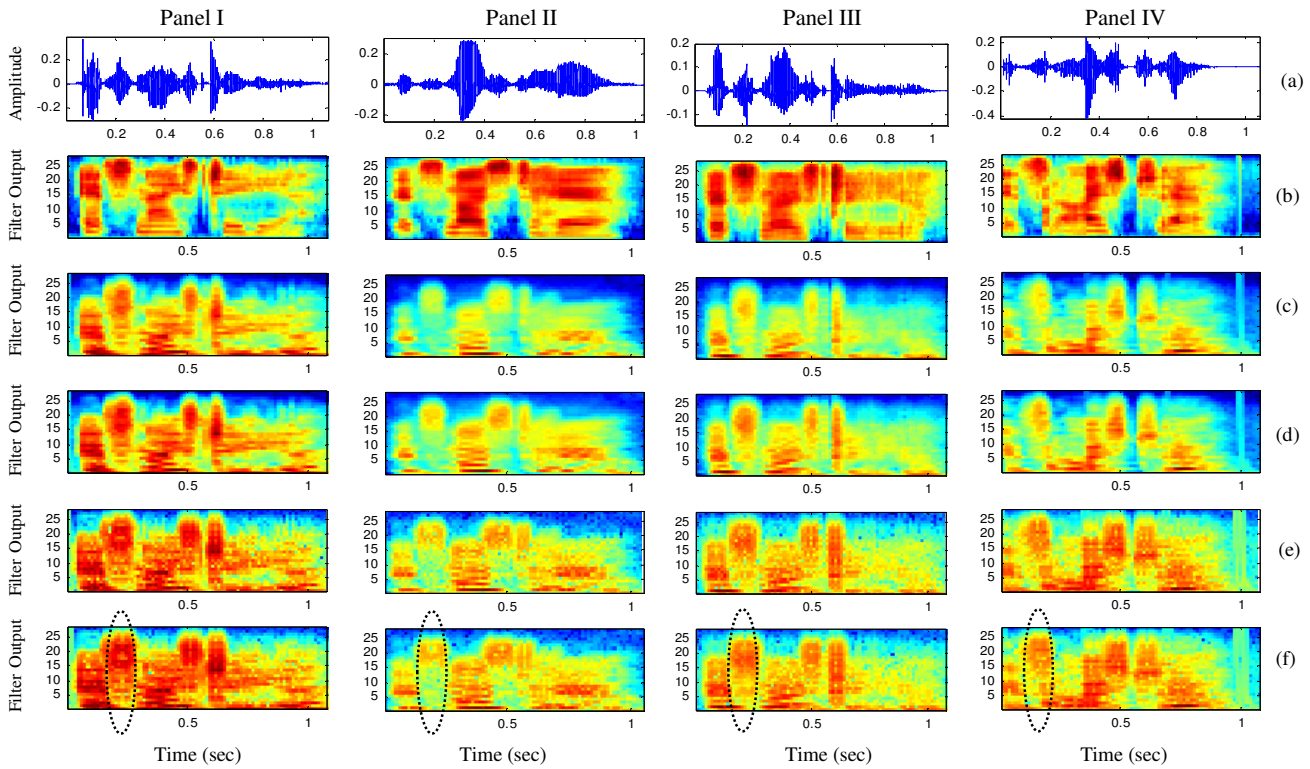


Fig. 4. Panel I: Natural speech, Panel II: vocoder-dependent synthetic speech (SS), Panel III: vocoder-dependent voice converted (VC) speech and Panel IV: vocoder-independent USS-based speech (SS MaryTTS): (a) speech signal waveform of the utterance /It's nice to hear/ from the SAS database [24], the subband energy representation in computation of (b) MFCC, (c) CFCC, (d) multiplication of subband envelope and average IF (without the derivative operation), the subband energy representation in computation of (e) CFCCIF (using one-point backward difference) [47] and (f) CFCCIFS (using symmetric difference operation). Dotted ovals show differences in natural and the spoofed speech.

IV. Spoofed Speech Detection (SSD) System

A. Parameterization

In this work, MFCC, CFCC, CFCCIF and CFCCIFS feature sets are used. These features are extracted from *25 ms* of frame with a shift of *50 %* between the frames. Both static (without 0^{th} energy coefficient) and dynamic features, i.e., delta (Δ) and delta-delta (ΔΔ) for all the feature sets are extracted. Thus, three different dimensions (*D*) of feature vector, i.e., *D1*: *12-D* static features, *D2*: *24-D* (*12*-static + *12*-Δ), *D3*: *36-D* (*12*-static + *12*-Δ + *12*-ΔΔ) are considered. To estimate the dynamic features, various analysis window intervals are considered for the derivative operation to know the best possible window size for SSD task. In addition, the individual contribution of the dynamic features is also analyzed. As discussed in Section III, the filters are designed with $\alpha=3$ and $\beta=0.035$ through intensive experiments. These values of α and β give a narrow shape and good quality factor to the auditory filters which help in IF estimation and then use change across frames to obtain the spoof-specific features.

B. Details of Database

The database provided for the ASVspoof 2015 challenge is used for this study. Details of the spoofing algorithms (S) are provided in [25]. The training and development dataset consisted of spoofed utterance generated by five spoofing algorithms (*S1*–*S5*) while evaluation data was based on *S1*–*S10*, i.e., both known and previously unseen (i.e., unknown) attacks. The *S3*, *S4* and *S10* are SS spoof and remaining are VC spoofs. Spoofing algorithm *S5* uses Mel Log Spectrum Approximation (MLSA) filter [59] and *S10* is implemented with USS-based open-source Modular Architecture for Research on speech sYnthesis (MARY) TTS system [60] that uses FESTIVAL framework [61] for speech synthesis. Remaining spoofs were generated by Speech Transformation and Representation using Adaptive Interpolation Weighted Spectrum (STRAIGHT) vocoder [62].

C. Model Training and Score-Level Fusion

In this study, we use a binary GMM-based classifier with *128* mixtures for modeling the classes corresponding to natural and spoofed speech. The GMM models are built on the training set provided for the ASVspoof 2015 challenge. The GMM for natural speech (λ_{nat}) is built using genuine (i.e., natural) utterances while GMM for spoofed speech (λ_{syn}) is built with spoofed utterances. Final scores on a test sequence *Y* are represented in terms of log-likelihood ratio (*LLR*) obtained from the likelihood value of natural and synthesized speech model. The decision of the test speech being human or spoofed is based on the *LLR*, i.e.,

$$LLR = \log\ (p\ (Y \mid \lambda_{nat})) - \log\ (p\ (Y \mid \lambda_{syn})), \qquad (18)$$

where $p(Y \mid \lambda_{\text{nat}})$ and $p(Y \mid \lambda_{\text{syn}})$ are the likelihood scores from the GMM for human speech and spoofed speech, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2016.2647201, IEEE Journal of Selected Topics in Signal Processing

To utilize possible complementary information captured by the feature sets, their score-level fusion is used, i.e.,

$$LLk_{combine} = (1-\alpha_f) LLk_{MFCC} + \alpha_f LLk_{feature2}, \quad (19)$$

where $LLk_{combine}$ is the combined log-likelihood score of two scores MFCC and *feature2* (i.e., either CFCC or CFCCIF or CFCCIFS), respectively. The weights of the scores are decided by fusion parameter α_f and are optimized *w.r.t* performance of system. We consider score-level fusion to know the contribution of the individual set of features and to avoid the higher feature dimension due to feature-level fusion.

D. Performance Measures

In evaluating a binary classifier, two types of errors exist, namely, False Acceptance (FA) and False Rejection (FR). A stand-alone detector system could falsely reject a genuine trial (a false rejection) to the ASV system or falsely accept a spoof or impostor trial and allow it to pass through an ASV system. The error rates are expressed as False Acceptance Rate (FAR), i.e., ratio of FA to actual number of positives (natural) and False Rejection Rate (FRR), i.e., ratio of FR to actual number of negatives (spoofed). Based on the FRR and FAR, the DET curve is used to measure the performance of various features [63]. It gives uniform treatment to both FRR and FAR for evaluation of system performance. In the DET curve, the operating point where FAR and FRR becomes equal is referred to as EER. Thus, EER is used as one of the performance measure [64]. Here, while testing, the average % EER is estimated by considering the natural speeches as positive class and all spoofed speeches from the various spoofing algorithms as the negative class. The threshold th_{EER} at the % EER serves as a boundary between output of positive and negative classes.

V. Experimental Results on Development Set

The performance of SSD system is evaluated considering the effect of number of subband filters, pre-emphasis filter, the contribution of static and dynamic features and dependency of the available countermeasures on spoofing algorithms.

A. Choice of Number of Subband Filters

Fig. 5 shows the dependency of the % EER on the number of subbands for Mel filterbank and cochlear filterbank. It is observed that overall the % EER decreases from *D1* to *D3* feature vector. The MFCC feature set has high % EER for less number of subband filters as compared to CFCC, CFCCIF and CFCCIFS. The % EER of the features do not vary much after *25* subband filters especially for *D3* feature vector. Hence, instead of using a large number of subband filters, we use slightly greater than *25,* i.e., *28* subband filters for all feature sets used in the remaining set of experiments.

B. Effectiveness of Pre-emphasis on Speech Signal

To study the dependence of the features on pre-emphasizing the speech signal, the % EER was obtained with pre-emphasis (*P*) and using *no* pre-emphasis (*nP*) for MFCC, CFCC, CFCCIF and CFCCIFS features sets. As shown in Fig. 6, MFCC features have a sensitive dependence to pre-emphasis, i.e., for *nP*, its % EER increases significantly for all sets of feature vectors. On the other hand, the % EER of CFCC-based features (with *P* or *nP*) are almost constant for all feature vectors. In fact, on an average, CFCC, CFCCIF and CFCCIFS feature sets perform better without explicit pre-emphasis. Thus, the performance of the cochlear filter-based features is not significantly dependent on pre-emphasis. This is due to the embedded or inherent bandpass filtering (i.e., due to *admissibility* condition of cochlear filter, i.e., mother wavelet function $\psi(t)$ as in eq. (6) [47]). Thus, for all the experiments, MFCC is used with pre-emphasis filter (i.e., $1\text{-}0.97z^{-1}$) and cochlear filter-based features are used without explicit use of pre-emphasis filter.

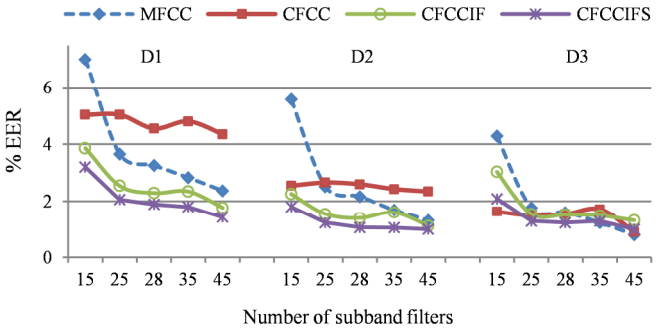
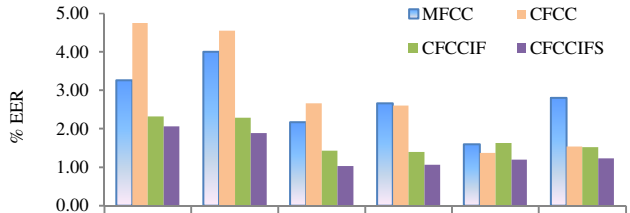


Fig. 5. Effect of various number of subband filters on the % EER for *D1*, *D2* and *D3* feature vectors for MFCC, CFCC, CFCCIF and CFCCIFS features.



	D1_P	D1_nP	D2_P	D2_nP	D3_P	D3_nP
MFCC	3.26	4.00	2.17	2.66	1.60	2.80
CFCC	4.75	4.55	2.66	2.60	1.37	1.54
CFCCIF	2.32	2.29	1.43	1.40	1.63	1.52
CFCCIFS	2.06	1.89	1.03	1.06	1.20	1.23

Fig. 6. Effect of pre-emphasis on % EER, using MFCC, CFCC, CFCCIF and CFCCIFS features (*P*=pre-emphasis with a pre-emphasis factor of *0.*97 and *nP*=no pre-emphasis on speech signal).

C. Effectiveness of Derivative Operation in CFCCIF

As discussed in Section III-D, the derivative operation (shown by dashed block in Fig. 3) before taking log and DCT indeed facilitates the SSD task. As seen in Table I, the % EER is less *with* derivative operation. In fact, using derivative in time-domain gives less % EER with just *D2* (static+Δ) feature vector. For higher-dimensional feature vector (*D3*), no difference is found in % EER with and without derivative. The CFCCIF feature set performs better with *D2* feature vector. However, to maintain consistency across all the feature sets, henceforth, we use *D3* feature vector in this paper.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2016.2647201, IEEE Journal of Selected Topics in Signal Processing

TABLE I
EER (IN %) OF CFCCIF FEATURE SET WITH AND WITHOUT DERIVATIVE

	D1	D2	D3
Without derivative	4.318	2.4878	1.5156
With derivative	2.287	1.4012	1.5156

D. *Effectiveness of Static and Dynamic Features*

To evaluate the performance of the discriminative features, the % EER is computed considering static, Δ and ΔΔ features. Table II shows the individual contribution for *12*-dimensional (*D1*) static features (excluding the 0^{th} coefficient) on the development set. It is seen that the MFCC features perform better than CFCC features (which is generally the case for the clean speech signal [46]). The CFCCIF features proposed in [47] performs better than CFCC (in terms of decrease in % EER). In fact, the CFCCIFS that use symmetric difference (for computing change across the frames) gave much less EER of *1.89* % with static features.

TABLE II
EER (IN %) OF *12-D* STATIC (S) FEATURE VECTOR

Features →	MFCC	CFCC	CFCCIF	CFCCIFS
12-static	3.26	4.55	2.29	1.89

A recent work in [36], has shown that the dynamic features alone can contribute effectively to the detection process and achieve almost similar or less % EER than the static features. Along the similar lines, Table III shows individual contribution for *12-D* Δ and *12-D* ΔΔ features for all the feature sets considered in the present work.

TABLE III
EER (IN %) OF *12-D* DYNAMIC FEATURES SETS USED ALONE

Frames	MFCC		CFCC		CFCCIF		CFCCIFS	
	Δ	ΔΔ	Δ	ΔΔ	Δ	ΔΔ	Δ	ΔΔ
w1	**3.77**	**4.69**	**1.54**	**0.77**	**2.69**	5.78	2.37	5.49
w2	5.83	6.23	5.18	3.75	2.83	**3.17**	**2.23**	**2.80**
w3	7.09	7.35	8.98	7.58	4.38	4.23	3.03	2.86
w4	8.15	8.49	11.41	10.21	7.06	5.78	4.58	4.03
Average	6.21	6.69	6.78	5.58	4.24	4.7	3.05	3.80

The dynamic features are estimated for four analysis frames ($2n_0+1$), i.e., with $n_0=1$ (*w1*), with $n_0=2$ (*w2*), with $n_0=3$ (*w3*) and with $n_0=4$ (*w4*) corresponding to *18.75 ms*, *31.25 ms*, *43. 75 ms* and *68.25 ms*, respectively, for a sampling frequency of *16 kHz*. It is observed that the dynamic features alone are effective for MFCC and CFCC only when a *w1* frame window is used. Likewise, CFCCIF and CFCCIFS perform well for the *w2* frame window. For larger frame window (i.e., *w3* and *w4*), the % EER increases significantly. This change is rather more for MFCC and CFCC features than that of the CFCCIF and CFCCIFS features as observed from the average values.

Next, the case of combined effect of the Δ and ΔΔ features are considered. It is observed from Table IV that with the Δ + ΔΔ features used together, the % EER improves rather than using dynamic features alone. The complementary information in Δ+ΔΔ features were added when the feature-level fusion is used. For the cochlear filter-based features, namely, CFCC, CFCCIF and CFCCIFS features, the performance was even better than the static features. From Table III, it was observed that the % EER for only ΔΔ features for CFCCIF and CFCCIFS features were more than *5* % with a *w1* frame window. However, when combined with their Δ features, the % EER went down to *1.77* % and *1.40* % for CFCCIF and CFCCIFS, respectively. Table IV shows that all the features perform better with a *w1* window and across all the window lengths, the CFCCIFS performs the best. By increasing the analysis window, the % EER seemed to be increasing for all the features. This increase in % EER was more for MFCC and CFCC as evident from the average over the windows. Thus, it is observed that the performance of MFCC and CFCC feature sets are more dependent on the window size as compared to the CFCCIF and CFCCIFS feature sets.

TABLE IV
EER (IN %) OF *24-D* (Δ+ΔΔ) DYNAMIC FEATURES

Frames	MFCC	CFCC	CFCCIF	CFCCIFS
w1	**3.09**	**1.00**	**1.77**	**1.40**
w2	5.72	4.55	2.12	1.54
w3	7.15	8.15	3.63	2.40
w4	8.84	10.69	5.63	3.66
Average	6.20	6.10	3.29	**2.25**

From Table IV, it can be concluded that *w1* is the best window for all the features considered in this paper to obtain relatively least % EER. Using this case, the combined effect of static, Δ and ΔΔ features are studied as shown by shaded cells in Table V. It is observed that with the *36-D*, (i.e., static+Δ+ΔΔ features), the % EER of MFCC reduced to *1.6* % as compared to using only static or only dynamic features alone. The CFCCIF and CFCCIFS features gave least % EER with *D2* feature vector due to reasons discussed in Section V-C.

TABLE V
EER (IN %) OF SCORE-LEVEL FUSION OF MFCC WITH CFCC (OR CFCCIF OR CFCCIFS) USING *D1*, *D2* AND *D3* FEATURE VECTORS AT VARIOUS α_f.

	MFCC+CFCC			MFCC+CFCCIF			MFCC+CFCCIFS		
α_f	D1	D2	D3	D1	D2	D3	D1	D2	D3
0	3.26	2.17	1.60	3.26	2.17	1.60	3.26	2.17	1.60
0.1	2.86	1.83	1.32	2.72	1.83	1.37	2.69	1.74	1.29
0.2	2.66	1.54	1.14	2.40	1.46	1.14	2.29	1.37	1.06
0.3	2.52	1.40	0.97	2.03	1.23	1.00	1.89	1.09	0.92
0.4	**2.43**	**1.32**	**0.89**	1.77	1.03	0.86	1.60	0.89	0.77
0.5	2.57	1.32	0.89	1.60	0.97	0.83	1.43	0.80	0.66
0.6	2.72	1.46	0.92	**1.52**	**0.89**	**0.83**	**1.37**	**0.71**	**0.66**
0.7	3.03	1.63	1.00	1.57	0.89	0.92	1.37	0.74	0.71
0.8	3.55	1.89	1.17	1.72	0.97	1.03	1.46	0.77	0.80
0.9	3.97	2.23	1.34	1.92	1.14	1.17	1.60	0.92	0.92
1	4.55	2.60	1.54	2.29	1.40	1.52	1.89	1.06	1.23

α_f = weight of score-level fusion, as per eq. (19)

Lastly, fusion of MFCC with either CFCC or CFCCIF or with CFCCIFS is considered. Table V shows that, the best % EER on the development set is obtained with a fusion weight, $\alpha_f = 0.4$ for CFCC and $\alpha_f = 0.6$ for CFCCIF and CFCCIFS. Thus, CFCCIF and CFCCIFS features on score-level fusion added more *complementary* information (than the MFCC alone) in reducing the % EER. The fusion of proposed CFCCIFS and MFCC using *D3* feature vector gave the least % EER of *0.66* amongst all the combinations (as shown by the dotted cells in Table V). In fact, scores obtained from MFCC and CFCCIF with $\alpha_f = 0.6$, i.e., with an EER of *0.83* % was

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2016.2647201, IEEE Journal of Selected Topics in Signal Processing

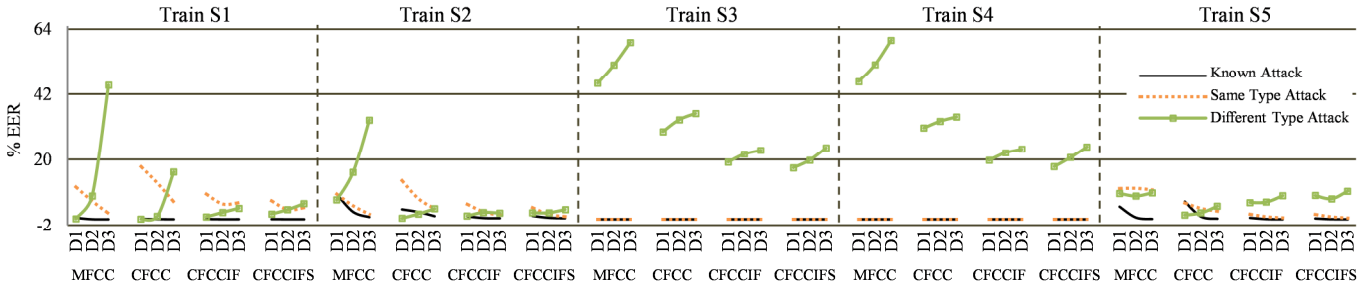


Fig. 7. Effect on % EER for known, same type and different types of attacks when trained with individual spoofs *S1*, *S2*, *S3*, *S4* and *S5* using MFCC, CFCC, CFCCIF and CFCCIFS feature sets with three different dimensions of feature vectors (i.e., *D1*, *D2*, *D3)* and tested on the entire development dataset.

submitted at the ASVspoof 2015 challenge which was found to be relatively the best performing system among all the *16* submissions [25]. Thus, the proposed MFCC-CFCCIFS system performs much better detection than the MFCC-CFCCIF-based SSD system.

E. Dependency on Spoofing Algorithms

To check the discriminative property of the proposed feature set in terms of the dependency of the spoofing algorithm, the systems were trained on individual spoofs and tested on all the spoofs of the development set. The development set consists of SS and VC spoofs which are further generated by different algorithms. Instead of considering the known and unknown attacks, we further break unknown attacks into two categories of 'same type' and 'different type' based on the method of generating speech (i.e., SS or VC). For example, for training with *S1* VC spoof: testing with speech from *S1* algorithm itself is 'known', testing with speech from another VC-based algorithm (i.e., *S2* and *S5*) is 'same type' and testing with speech from any SS-based spoofing algorithm (i.e., *S3* and *S4*) is 'different type'. Average of same type and different type constitutes 'unknown' attacks. From Fig. 7, the following observations can be made:

- For the known attacks: Each of the features works well for known attacks (shown by black solid line). The SS spoofs *S3* and *S4* obtained *0.0* % EER is obtained when tested with itself. For VC spoof, when tested by itself, the % EER decreased with MFCC, CFCC, CFCCIF and CFCCIFS feature sets with *D3* feature vector.
- For the same type of attack: The SS spoofs (*S3* and *S4*) performed the best to detect each other. However, training with VC-based *S1* and *S2* that uses STRAIGHT vocoder identified VC-based *S5* spoof generated using MLSA vocoder with an average *10.7* % and *8.7* % EER and detected each other with *3.4* % (*S2*) and *0.12* % (*S1*) EER, respectively. On the other hand, *S5* spoof detected *S1* with *2.4* % and *S2* with *5.53* % EER.
- For the different type of attack: The STRAIGHT-based *S1* or *S2* VC speech when used for training, detected STRAIGHT-based *S3* and *S4* quite well with CFCCIF and CFCCIFS features. However, when only *S5* VC spoof is used for training, it could not detect SS spoof well. Likewise, the SS spoof, when tested with VC spoof, gave very large % EER. For MFCC, around *50* % EER is observed which decreases to around *20* % for CFCCIFS features. The % EER increases for *D3* feature vector on testing with different type of spoof, especially for MFCC. On the whole, the trend decreased in % EER from MFCC to CFCCIFS features.

It was observed that training with VC spoof detected SS spoof to some extent. However, SS trained models could not detect VC spoofs (this is also indicative in Table VI). Fig. 8 shows the true (dotted) and false (solid) scores distribution of the testing data when trained only with *S1* VC spoof (top panel) and *S3* SS spoof (bottom panel). The scores are shown for *D1* features vector of MFCC, CFCC, CFCCIF and CFCCIFS. While training with VC spoof alone, features for synthetic speech could probably be captured and hence, the SS were detected in the testing phase which is evident from the single distribution of the false scores. Therefore, as models trained on VC detected both SS and VC speech, there is a single distribution of false scores. However, the models trained on SS could not detect VC speech, resulting in an M-like distribution of false scores. The part of the M-shaped distribution that is overlapping with the true scores (dotted line) is likely due to scores from the VC speech used in testing. The SS spoof model had been trained with features specific to it and hence, could not detect VC speech resulting in large overlapping regions leading to increased % EER.

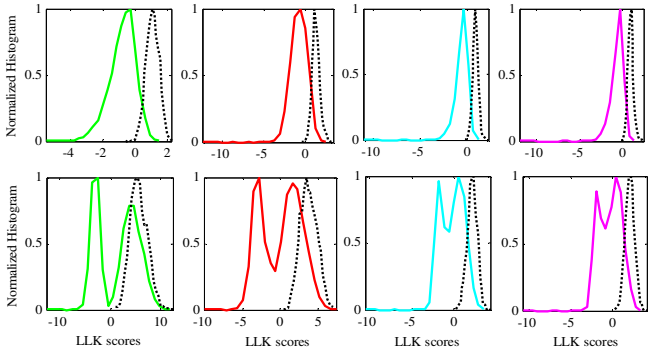


Fig. 8. The distribution for true scores (dotted line) and false scores (solid line) for *12-D* static features extracted from MFCC (green), CFCC (red), CFCCIF (cyan) and CFCCIFS (magenta) when trained with *S1* VC spoof (top panel) and *S3* SS spoof (bottom panel). The false scores show a single distribution when trained with VC spoof (top panel) and M-shaped distribution when trained SS spoof (bottom panel).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2016.2647201, IEEE Journal of Selected Topics in Signal Processing

Amongst all the feature sets considered here, CFCCIFS had the least overlap between true and false score distribution resulting in low % EER and better performance. Table VI shows % EER of known and unknown attacks when trained with individual spoofs and tested on the development set. The VC spoofs gave less EER for unknown attacks as it detected SS spoof to some extent. On the other hand, SS spoof gave very high % EER, i.e., > *30* % with MFCC and > *10* % with CFCCIFS. The shaded cells in Table VI show the best performance of known (Kn) and unknown attacks (Ukn).

TABLE VI
EER (IN %) FOR KNOWN (KN) AND UNKNOWN ATTACKS (UKN) WHEN TRAINED ON INDIVIDUAL SPOOFS AND TESTED ON ENTIRE DEVELOPMENT SET

Train →		MFCC		CFCC		CFCCIF		CFCCIFS	
Test		Kn	Ukn	Kn	Ukn	Kn	Ukn	Kn	Ukn
S1 (VC)	D1	0.40	5.48	0.17	8.8	0.08	4.62	0.05	3.97
	D2	0.03	6.96	0.06	6.57	0.02	3.71	0.02	3.25
	D3	0.00	23.48	0.01	10.83	0.03	4.63	0.01	4.59
S2 (VC)	D1	8.39	7.45	3.31	6.63	0.98	3.08	1.14	3.04
	D2	2.49	10.11	2.45	4.34	0.43	2.41	0.51	1.90
	D3	0.76	17.51	1.06	3.56	0.36	1.74	0.36	2.10
S3 (SS)	D1	0.00	34.21	0.00	22.06	0.00	14.35	0.00	12.93
	D2	0.00	38.93	0.00	25.06	0.00	16.37	0.00	14.83
	D3	0.00	44.65	0.00	26.61	0.00	17.32	0.00	17.94
S4 (SS)	D1	0.00	34.65	0.00	22.95	0.00	14.8	0.00	13.19
	D2	0.00	38.97	0.00	24.66	0.00	16.69	0.00	15.51
	D3	0.00	45.19	0.00	25.74	0.00	17.62	0.00	18.23
S5 (VC)	D1	4.22	9.46	5.91	3.43	0.49	3.66	0.28	4.79
	D2	0.60	9.06	0.89	2.89	0.08	3.30	0.04	3.83
	D3	0.12	9.34	0.21	3.56	0.01	4.19	0.03	4.92

VI. EXPERIMENTAL RESULTS ON EVALUATION SET

On the development set, it was observed that instead of using static or dynamic features alone, their combination (i.e., *D3* feature vector) gives less % EER. In addition, score-level fusion of MFCC and cochlear filter-based features one at a time, i.e., MFCC with CFCC (or CFCCIF or CFCCIFS) features gave the least % EER. Based on the results of the development set, parameterization is chosen for evaluation set.

A. Results of Score-Level Fusion

Table VII shows the results in % EER on the evaluation data after fusion of MFCC features with CFCC (or CFCCIF or CFCCIFS) sets using *D1*, *D2* and *D3* feature vectors.

TABLE VII
EER (IN %) OF SCORE-LEVEL FUSION OF MFCC WITH CFCC (OR CFCCIF OR CFCCIFS) USING *D1*, *D2* AND *D3* FEATURE VECTORS FOR VARIOUS FUSION FACTORS (α_f) ON EVALUATION DATASET

	MFCC+CFCC			MFCC+CFCCIF			MFCC+CFCCIFS		
α_f	D1	D2	D3	D1	D2	D3	D1	D2	D3
0.0	5.50	5.49	4.26	5.50	5.49	4.26	5.50	5.49	4.26
0.1	4.86	4.23	3.19	4.92	4.52	3.46	4.84	4.41	3.36
0.2	4.49	3.67	2.67	4.46	3.96	2.99	4.31	3.74	2.79
0.3	4.21	3.19	2.26	4.04	3.53	2.66	3.84	3.24	2.41
0.4	4.09	2.81	1.99	3.74	3.13	2.39	3.46	2.81	2.06
0.5	4.03	2.58	1.79	3.51	2.79	2.19	3.17	2.45	1.80
0.6	**4.02**	2.45	1.69	3.35	2.55	2.03	2.98	2.13	1.60
0.7	4.15	**2.40**	1.62	3.23	2.38	1.91	2.85	1.91	1.49
0.8	4.33	2.45	**1.61**	**3.16**	2.30	**1.87**	**2.74**	1.80	**1.45**
0.9	4.64	2.58	1.63	3.19	**2.28**	1.91	2.74	**1.75**	1.48
1.0	4.98	2.78	1.74	3.37	2.34	2.07	2.81	1.81	1.60

α_f = weight of score-level fusion, as per eq. (19)

Table VII shows that score-level fusion at *D3* feature vector gives best results for all the features. For the development set, the best % EER was obtained with $\alpha_f = 0.4$ for CFCC and with $\alpha_f = 0.6$ for both CFCCIF and CFCCIFS feature sets. However, for the evaluation set, the fusion factor changes to $\alpha_f = 0.8$, due to the presence of unknown attacks. With $\alpha_f = 0.8$, the cochlear-based features contribute more in reducing the % EER for unknown attacks. In fact, the *1.6* % EER of the *D3* feature vector for CFCCIFS alone is almost equal to the least EER of *1.45* % after fusion.

B. Spoof Dependency

The attack-dependent % EER for the individual spoofs of the evaluation set using all the four feature vectors (i.e., MFCC, CFCC, CFCCIF and CFCCIFS) are shown in Table VIII. It is observed that when both the VC and SS spoofs are used for training, the known and unknown attacks are detected quite well (except *S10*). For *D3* feature vector, MFCC features achieved least *0.37* % EER for known attacks. However, it achieved very high *40* % EER for *S10* which increased the average EER to *4.26* %. The CFCCIFS feature set gave an EER of *0.45* % for known attacks and least EER of *2.73* % on unknown attacks with *11.7* % EER for *S10* spoof. The CFCC feature set also obtained less % EER for *S10* spoof. However, its % EER for other spoofs was more than that of CFCCIF and CFCCIFS. Amongst known attacks, *S2* and *S5* spoof were found to be difficult to detect and for unknown attacks, the vocoder-independent *S10* spoof was found to be toughest, followed by *S6*, *S9*, *S8* and *S7*. Overall, the CFCCIFS feature set works quite well for known attacks and detected unknown spoofed speech even if a similar type was not available while training.

Fig. 9 shows the % EER for known and unknown attacks with *D3* feature vector at various fusion factors when MFCC is fused with either CFCC, or CFCCIF or CFCCIFS as in eq. (19). While MFCC features, when used alone, gave least % EER for known attacks, the cochlear filter-based features gave least % EER for unknown attacks even without fusion with MFCC. Thus, some contribution of MFCC is required for best performance with known attacks.

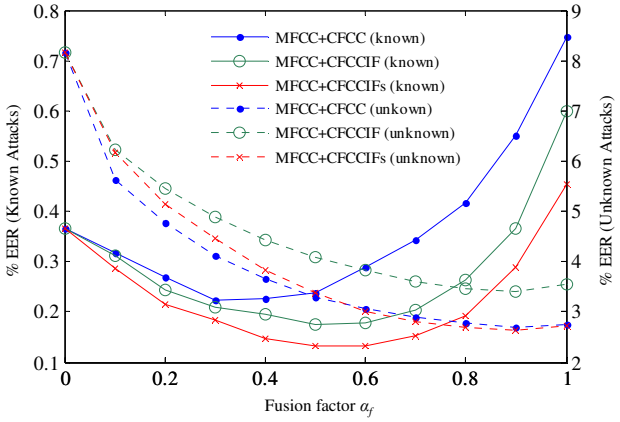


Fig. 9. The % EER of known attacks (solid line) and unknown attacks (dashed line) for *D3* feature vector for score-level fusion of MFCC with CFCC (blue), CFCCIF (green) and CFCCIFS (red) (as per eq. (19)).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2016.2647201, IEEE Journal of Selected Topics in Signal Processing

TABLE VIII
EER (IN %) FOR KNOWN AND UNKNOWN ATTACKS USING MFCC, CFCC, CFCCIF AND CFCCIFS FOR ALL FEATURE VECTORS AND % EER OBTAINED ON FUSION OF MFCC WITH CFCC (OR CFCCIF OR CFCCIFS) USING OPTIMUM VALUE OF α_f FOR *D3* FEATURE VECTOR ON THE EVALUATION DATASET

Feature vector		Known					Unknown					Average		
	Features	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Known	Unknown	Average
D1 (static)	MFCC	0.11	3.78	0.00	0.00	4.51	3.64	0.40	0.03	1.01	41.47	1.68	9.31	5.49
	CFCC	0.14	2.62	0.00	0.00	11.33	4.77	0.43	0.13	1.20	29.15	2.82	7.13	4.98
	CFCCIF	0.02	0.49	0.00	0.00	3.33	1.28	0.09	0.18	0.17	28.15	0.77	5.97	3.37
	CFCCIFS	0.03	0.46	0.00	0.00	2.59	0.99	0.09	0.23	0.16	23.52	**0.61**	**5.00**	**2.81**
D2 (static+Δ)	MFCC	0.02	1.58	0.00	0.00	1.63	1.75	0.10	0.00	0.00	49.58	0.65	10.33	5.49
	CFCC	0.10	2.53	0.00	0.00	4.98	2.51	0.25	0.08	0.55	16.76	1.52	4.03	2.78
	CFCCIF	0.01	0.22	0.00	0.00	1.49	0.65	0.07	0.40	0.05	20.50	0.35	4.33	2.34
	CFCCIFS	0.03	0.23	0.00	0.00	1.26	0.45	0.08	0.54	0.09	15.43	**0.30**	**3.32**	**1.81**
D3 (static+Δ+ΔΔ)	MFCC	0.01	0.99	0.00	0.00	0.83	0.90	0.05	0.00	0.00	39.72	**0.37**	8.15	4.26
	CFCC	0.04	1.39	0.00	0.00	2.30	1.04	0.12	0.06	0.21	12.28	0.75	2.74	1.74
	CFCCIF	0.03	0.72	0.00	0.00	2.24	0.98	0.16	0.88	0.29	15.42	0.60	3.55	2.07
	CFCCIFS	0.03	0.50	0.00	0.00	1.74	0.71	0.14	0.96	0.16	11.71	0.45	**2.73**	**1.60**
D3 With optimum α_f=*0.8*	MFCC+CFCC	0.016	0.74	0.00	0.00	1.33	0.68	0.076	0.00	0.12	13.08	0.42	2.79	1.605
	MFCC+CFCCIF	0.00	0.36	0.00	0.00	0.97	0.5	0.043	0.082	0.049	16.72	0.27	3.48	1.872
	MFCC+CFCCIFS	0.00	0.24	0.00	0.00	0.72	0.31	0.033	0.098	0.038	13.03	0.18	2.70	1.446

α_f = weight of score-level fusion, as per eq. (19)

Table VIII shows the % EER for optimum α_f of *0.8* for CFCC, CFCCIF and CFCCIFS. Amongst all possible score-level fusions, MFCC+CFCCIFS combination is found to perform relatively the best.

The DET curves for MFCC, CFCC, CFCCIF and CFCCIFS feature sets when used alone are shown in Fig. 10 (a). It is observed that the FRR of MFCC was very high for a given FAR which is not suitable for ASV systems. From MFCC to CFCC, there is a significant decrease in FRR (as shown in Fig. 10 (a) by dotted oval) which further reduces for CFCCIFS. Fig. 10 (b) shows DET curves for cochlear filer features after fusion with MFCC for $\alpha_f = 0.8$. A clear improvement in both FRR and FAR is observed with MFCC+CFCCIFS than with CFCC and CFCCIF. The fusion of MFCC+CFCCIF with $\alpha_f = 0.6$ was submitted at the ASVspoof 2015 challenge (as decided from the results on development set). However, in this paper, MFCC+CFCCIFS gives better % EER and the best performance amongst all the other fusion combinations.

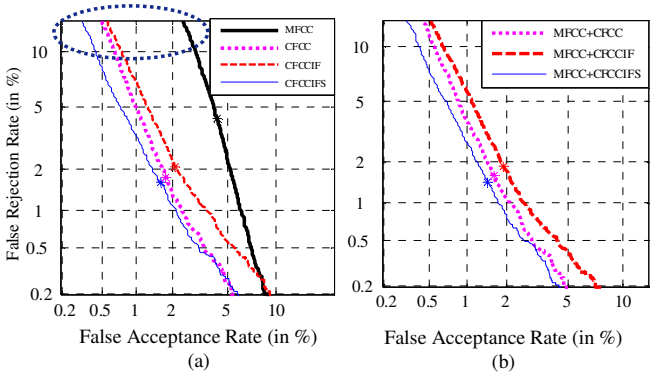


Fig. 10. The DET curves for (a) MFCC, CFCC, CFCCIF and CFCCIFS feature sets used alone (b) score-level fusion of MFCC + CFCC, MFCC + CFCCIF and MFCC + CFCCIFS with a factor $\alpha_f = 0.8$ (as per eq. (19)).

C. *Dependency on Spoofing Algorithms*

Just as in the case of development set, here, the testing of the entire evaluation data is carried out when trained on individual *S1*, *S2*, *S3*, *S4* and *S5* spoofs. It was observed in Section V-E, that the known and same types of spoofs were easily identified. However, for 'different type' of spoof on training with VC spoofs, the SS spoofs were identified. However, when trained on SS only, SS was detected well and not VC.

We continue the same analysis here, using 'same type' and 'different type' of spoof. However, we consider *S10* separately as it is non-vocoder type and its performance highly affected the average % EER. The interpretations for 'known type' remain similar as discussed for Fig. 7 (and hence, not shown here again). Fig. 11 shows that as in the development set, the trend is similar, i.e., for the 'same type' of spoofs, SS identified its same type with almost *0.00* % EER. VC spoof identified its same type quite well and the % EER decreased from MFCC to CFCCIFS features. On the other hand, for 'different type' of spoof, VC spoofs gave less % EER when tested on SS attacks as compared to SS spoof that gave very high % EER on testing with VC spoof. Overall, on training with *S1*, *S2* and *S5* (VC spoof), MFCC feature set gave an EER of *4.71 %, 4.57 %* and *2.12 %* and CFCCIFS gave an average EER of *1.43 %, 1.00 %* and *2.10 %* for all vocoder-based (*S1-S9*) spoofs averaged over all feature dimensions. This analysis was independent of MARY TTS (i.e., *S10*).

Considering *S10* separately, VC spoofs (*S1*, *S2* and *S5*) when tested with *S10* gave large % EER for MFCC features, i.e., the detection rate were around *20-70 %* with MFCC. The % EER gradually decreased to *15-25 %* when CFCCIFS features are used with *D3* feature vector. Interestingly, when trained on *S1* spoof with CFCCIFS using *D3* feature vector, as low as *2.6 %* EER is achieved. Similarly, on testing with SS trained models, *S10* gave as low as *3 %* with MFCC using *D1* feature set which increased to *10-50 %* when other features were used. For MFCC features, the % EER increases from *D1* to *D3* feature vector, while the pattern of % EER for cochlear filter-based features is found to be random. On listening to MARY TTS speech utterances from SAS database [24], they were found to be unintelligible [26]. Thus, it should be possible to identify this kind of spoof. However, the modeling needs to be done appropriately and techniques to deal with vocoder-independent speech needs to be explored further. On the whole, the proposed CFCCIFS feature set performed better due to use of auditory filterbank (than the triangular filterbank used in MFCC) and due to the concept of IF information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2016.2647201, IEEE Journal of Selected Topics in Signal Processing

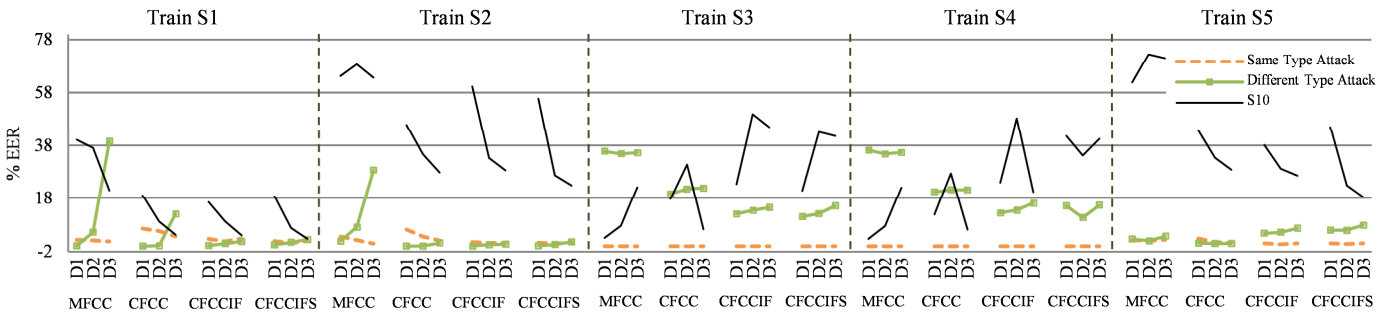


Fig. 11. Effect on % EER for same type, different types and *S10* attacks when trained with individual spoofs *S1*, *S2*, *S3*, *S4* and *S5* using MFCC, CFCC, CFCCIF and CFCCIFS feature sets with three different dimensions of vectors (i.e., *D1*, *D2*, *D3*) and tested on the entire evaluation dataset.

VII. Summary and Conclusions

In this study, cochlear filter-based features that use wavelet-like auditory transform and the variation of envelope and average IF features across subband at the output of cochlear filterbank are proposed. The performance of MFCC, CFCC, CFCCIF and CFCCIFS feature sets are compared. Use of symmetric difference to estimate the variations of subband energy representation in CFCCIFS brings out more prominent differences between natural and spoofed speech than MFCC, CFCC and CFCCIF. The cochlear filter-based features had an advantage that *no* explicit pre-emphasis was needed prior to feature extraction which reduces pre-processing requirements as compared to MFCC. Our study considered separately the use of static and dynamic features for all the feature sets. It was observed that both static and dynamic (Δ and ΔΔ) features jointly are needed to achieve low % EER. In addition, the dynamic (Δ and ΔΔ) information when estimated with a smaller analysis window was found to be more useful for the SSD task.

We have recognized the feasibility of the SSD systems to deal with both known and unknown attacks. It was observed that CFCCIFS feature set worked very well for the unknown attacks which is one of the most challenging subtask for the SSD system. Here, for the ASVspoof 2015 challenge database, the proposed CFCCIFS feature set gave quite low % EER for known attacks and relatively lowest % EER for unknown attacks. It has been observed that the fusion of MFCC and CFCCIFS gave least average % EER among the several combinations. In the literature, for SSD task, use of phase-based countermeasures has been commonly used. At the ASVspoof 2015 challenge, phase-based approaches were used because state-of-the-art SS and VC techniques use vocoder which lacks phase information. These countermeasures gave almost *0.00* % EER for known attacks. However, many of these approaches failed for non-vocoder MARY TTS spoof (*S10*) with almost *20-40* % EER. Our proposed CFCCIFS features gave as low as *11* % EER for MARY TTS spoof. In general, it was observed that amongst the SS and VC vocoder-based spoofs, VC spoofs were hard to detect than SS spoof. The vocoder-independent MARY TTS synthetic speech spoof was found to be toughest amongst all the spoofs.

The LFCC and CQCC feature sets gave better performance on *S10* spoof [36], [38]. With the use of only dynamic information and by using the EER computation as in [25], the LFCC and CQCC features obtain an average EER of *0.899* % and *0.255* % with *8.185* % and *1.065* % for the *S10* spoof, respectively. Using this framework, the average EER by the MFCC-CFCCIF ($\alpha_f = 0.6$) system was *1.211* % with *8.49* % for *S10* spoof [47], while the proposed MFCC-CFCCIFS system ($\alpha_f = 0.8$) gives an average EER is *0.922* % with *5.7* % for the *S10* spoof which is an improvement over the MFCC-CFCCIF system.

We have identified that the performance of the features could also be measured on the basis of how well the 'same type' and 'different type' of attacks are identified (as defined in Section V (E)). That is, we consider the case of availability of only one type of spoofing algorithm (either SS or VC) available for the training. We also found through intensive experiments that on training with VC spoofs, the SS spoofs were identified well. In any case, it was found that the proposed CFCCIFS feature set is very effective to test various spoofs (SS or VC) even during training with any available spoof as compared to MFCC, CFCC and CFCCIF feature sets. Thus, the present work uses different experimental setup and in fact, the present setup that uses of all the spoofs together to obtain a single average % EER is also challenging. For SSD task, flaws of vocoder are used to develop countermeasures. However, we have observed improvements when we explore features which are specific to the natural human speech production mechanism and that spoofed speech cannot mimic them so easily. Our related future research work will be directed towards exploring the potential of proposed CFCCIFS feature sets for language-independent SSD task and robustness under signal degradation or noisy conditions. In addition, as CFCC has been used in speaker identification and IF is known to have formant tracking application, therefore, we look forward to use of CFCCIF-based features in speaker recognition and related tasks.

Acknowledgments

The authors would like to thank Mr. Hardik B. Sailor for his kind help and co-operation during the various technical discussions in this research work. The authors also thank the authorities of DA-IICT, Gandhinagar for their kind support and cooperation in carrying out the research work.

REFERENCES

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. of the IEEE*, vol. 85, no. 9, pp. 1437-1462, Sept. 1997.

[2] A. E. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE,* vol. 64, no. 4, pp. 475-487, April 1976.

[3] K. Tokuda, H. Zen and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Workshop on Speech Synthesis*, Santa Monica, CA, USA, pp. 227-230, 2002.

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

[5] J. Yamagishi, T. Kobayashi, Y. Nakona, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 17, no.1, pp.66-83, Jan. 2009.

[6] Y. Stylianou, O. Cappe and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.,* vol. 6, no. 2, pp. 131-142, Mar. 1998.

[7] J.-F. Bonastre, D. Matrouf and C. Fredouille, "Transfer function-based voice transformation for speaker recognition," in *Proc. Odyssey 2006: The Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, 2006, pp. 1-6.

[8] Z. Wu, et al., "Spoofing and countermeasures for speaker verification: A survey," *Speech Comm.,* vol. 66, pp. 130-153, Feb. 2015.

[9] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech," in *Proc. Eur. Conf. Speech Process. Techno* .(*EUROSPEECH*), Budapest, Hungary, 1999, pp. 1223-1226.

[10] T. Masuko, K. Tokuda and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, Beijing, China, 2000, pp. 302-305.

[11] P. L. De Leon, M. Pucher and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Proc. Odyssey 2010: The Speaker and Language Recognition Workshop,* Brno, Czech Republic, 2010, pp. 151-158.

[12] 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 Audio, Speech, & Lang. Process.,* vol. 20, no. 8, pp. 2280-2290, Oct. 2012.

[13] D. Matrouf, J.-F. Bonastre and C. Fredouille, "Effect of speech transformation on impostor acceptance," in *Proc. Int. Conf. on Acous., Speech and Sig. Process.* (*ICASSP)*, Toulouse, France, 2006, pp. 933-936.

[14] J. F. Bonastre, D. Matrouf and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Proc. INTERSPEECH*, Antwerp, Belgium, 2007, pp. 2053-2056.

[15] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouche and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Lang. Process.,* vol. 19, no. 4, pp. 788-798, 2011.

[16] T. Kinnunen, et al., "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *Proc. IEEE Int. Conf. on Acous., Speech and Sig. Process.* (*ICASSP*), Kyoto, Japan, 2012, pp. 4401-4404.

[17] B. Steward, P. L. De Leon and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *Proc. INTERSPEECH,* Portland, Oregon, USA, *2012,* pp. 370-373.

[18] Z. Wu, X. Xiao, E. S. Chng and H. Li, "Synthetic speech detection using temporal modulation feature," in *Proc. Int. Conf. on Acous., Speech and Sig. Process.* (*ICASSP)*, Vancouver, BC, Canada, 2013, pp. 7234-7238.

[19] P. L. De Leon, I. Hernaez, I. Saratxaga, J. Yamagishi, and M. Pucher, "Detection of synthetic speech for the problem of imposture," in *Proc. Int. Conf. on Acous., Speech and Sig. Process.* (*ICASSP)*, Prague, Czech Republic, 2011, pp. 4844-4847.

[20] J. Sanchez, et al., "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Trans. Info. Forensics and Security,* vol. 10, no. 4, pp. 810-820, April 2015.

[21] Z. Wu, E. S. Chng and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *INTERSPEECH*, Portland, Oregon, USA, pp. 1700-1703.

[22] Z. Wu, T. Kinnunen, E. S. Chng, H. Li, and E. Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: The telephone speech case," in *Proc. Asia-Pacific Sig. & Infor. Process. Assoc. Annual Summit and Conf. (APSIPA ASC),* Hollywood, CA, USA, 2012, pp. 1-5.

[23] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and anti-spoofing in the i-vector space," *IEEE Trans. Info. Forensics and Security,* vol. 10, no. 4, pp. 821-832, Feb. 2015.

[24] Z. Wu, et al., "SAS: A speaker verification spoofing database containing diverse attacks," in *Proc. IEEE Int. Conf. on Acous., Speech and Sig. Process.* (*ICASSP*), Brisbane, 2015, pp. 4440-4444.

[25] Z. Wu, et al., "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge", in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2037-2041.

[26] Z. Wu, et al., "Anti-spoofing for text-independent speaker verification: An initial database, comparision of countermeasures, and human performance," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 24, no. 4, pp. 798-783, April 2016.

[27] Y. Liu, Y. Tian, L. He, J. Liu and M. T. Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2082-2086.

[28] X. Xiao, et al., "Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2052-2056.

[29] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas and D. Erro, "The AHOLAB RPS SSD spoofing challenge 2015 submission," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2042-2046.

[30] L. Wang, Y. Yoshida, Y. Kawakami and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2092-2096.

[31] I. Saratxaga, J. Sanchez, Z. Wu, I. Hernaez, and E. Navas, "Synthetic speech detection using phase information," *Speech Comm*., vol. 81, pp. 30-41, April 2016.

[32] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of CRIM system for the automatic speaker verification spoofing and countermeasure challenge 2015," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2072-2076.

[33] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2077-2081.

[34] H. Bhavsar, T. B. Patel, and H. A. Patil, "Novel nonlinear prediction based features for spoofed speech detection," in *Proc. INTERSPEECH*, San Francisco, USA, 2016, pp. 155-159.

[35] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik and V. Shchemelinin, "STC anti-spoofing systems for the ASVspoof 2015 challenge," in *Proc. IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP),* Shanghai, China, 2016, pp. 5475-5479.

[36] M. Sahidullah, T. Kinnunen and C. Hanilçi, "A comparison of features for synthetic speech detection," in Proc. *INTERSPEECH*, Dresden, Germany, 2015, pp. 2087-2091.

[37] H. Yu, et al., "Effect of multi-condition training and speech enhancement methods on spoofing detection," in *Int. Workshop on Sensing, Process. and Learning for Intelligent Machines (SPLINE),* Aalborg, Denmark, 2016, pp. 1-5.

[38] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients," in *Proc. Odyssey 2016: The Speaker and Language Recognition Workshop*, Bilbao, Spain, 2016, pp. 283-290.

[39] T. B. Patel and H. A. Patil, "Effectiveness of fundamental frequency (F_0) and strength of excitation (SoE) for spoofed speech detection," in

Proc. IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP), Shanghai, China, 2016, pp. 5105-5109.

[40] S. Weng, et al., "The SYSU system for the Interspeech 2015 automatic speaker verification spoofing and countermeasures challenge," Cornell University Library, arXiv:1507.06711, 2015.

[41] N. Chen, Y. Qian, H. Dinkel, B. Chen and K. Yu, "Robust deep feature for spoofing detection-The SJTU system for ASVspoof 2015 challenge," in *Proc. INTERSPEECH* , Dresden, Germany, 2015, pp. 2097-2101.

[42] M. H. Soni, T. B. Patel, and H. A. Patil, "Novel subband autoencoder features for detection of spoofed speech," *in INTERSPEECH*, San Francisco, USA, pp. 1820-1824.

[43] J. Villalba, A. Miguel, A. Ortega and E. Lleida, "Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2064-2071.

[44] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Pearson India, Eight Impression, 2012.

[45] Q. Li, "An auditory-based transform for audio signal processing," in *IEEE Workshop on Applications of Sign. Process. to Audio and Acous.*, New Paltz, NY, 2009, pp. 181-184.

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

[47] 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 *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2062-2066.

[48] N. Singh, N. Bhendwade, and H. A Patil, "Novel cochlear filter based cepstral coefficients for classification of unvoiced fricatives," *Int. J. on Natural Lang. Computing*, vol. 3, no. 4, pp. 21-40, Aug. 2014.

[49] J. H. L. Hansen and D. T. Chappell, "An auditory-based distortion measure with application to concatenative speech synthesis," *IEEE Trans. Speech Audio Process.,* vol. 6, no. 5, pp. 489-495, 1998.

[50] K. K. Paliwal and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Comm.*, vol. 45, no. 2, pp. 153-170, Feb. 2005.

[51] H. Yin, V. Hohmann, and C. Nadeu, "Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency," *Speech Comm.*, vol. 53, no. 5, pp. 707-715, May 2011.

[52] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," in *Proc. Eur. Conf. Speech Process. Techno. (EUROSPEECH)*, Aalborg, Denmark, 2001, pp. 759-762.

[53] L-W. Chen, W. Guo, and L-R.Dai, "Speaker verification against synthetic speech," in *Proc. Int. Symposium on Chinese Spoken Lang. Process. (ISCSLP)*, Sun Moon Lake, Taiwan, 2010, pp. 309-312.

[54] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," in *Proc. IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Shnaghai, China, 2016, pp. 2119-2123.

[55] S. G. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed., Academic Press, 1998.

[56] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acous. Soc. Amer. (JASA)*, vol. 118, no. 2, pp. 887-906, Aug. 2005.

[57] S. Shamma and D. Klein, "The case of the missing pitch templates: How harmonic templates emerge in the early auditory system," *J. Acous. Soc. Amer. (JASA),* vol. 107, no. 5, pp. 2631-2644, 2000.

[58] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Letters to Nature,* vol. 416, no. 6876, pp. 87-90, 2002.

[59] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. IEEE Int. Conf. on Acous., Speech and Sig. Process.* (*ICASSP*), San Francisco, CA, USA, 1992, pp. 137-140.

[60] "The MARY Text-to-Speech System (MaryTTS)," [Available Online]: http://mary.dfki.de/ {Last accessed: 24th August 2015}.

[61] A. Black, P. Taylor and R. Caley, "The Festival speech synthesis system," 1988. [Available Online]: http://festvox.org/festival/ {Last accesssed: 24th January 2016}.

[62] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds," *Speech Comm.,* vol. 27, no. 3-4, pp. 187-207, April 1999.

[63] A. Martin, G. Doddington, T. Kamm, and M. Ordowski, "The DET curve in assessment of detection task performance," in *Proc. Eur. Conf. Speech Process. Techno. (EUROSPEECH)*, Greece, 1997, pp. 1895-1898.

[64] DET-Curve Plotting software for use with MATLAB. [Available Online].http://www.itl.nist.gov/iad/mig/tools/DETware_v2.1.targz.htm, {Last Accessed 09th August 2016}.

Tanvina B. Patel received her B.E. degree in E.C. Engg. from Govt. Engg. College (GEC), Surat, in 2009 and the M.E. degree in Comm. Systems Engg. (C.S.E) from L. D. College of Engg., Ahmedabad, Gujarat, in 2012. Currently, she is a Ph.D. scholar at DA-IICT, Gandhinagar, India, and research assistant at DeitY, (Govt. of India) sponsored consortium project on Automatic Speech Recognition (ASR) for Agricultural Commodities in Indian Languages-Phase II. She was also associated with the consortium project on Development of Text-to-Speech (TTS) Systems in Indian Languages-Phase-II. Her research interest includes speech signal analysis, speech synthesis, and spoof speech detection for voice biometrics.

Ms. Patel is a student member of IEEE, IEEE SPS and International Speech Communication Association (ISCA). She has received ISCA, Microsoft Research (MSR) India and IEEE SPS travel grant to present research papers at INTERSPEECH 2013, INTERSPEECH 2014 and ICASSP 2016, respectively.

Hemant A. Patil received his B.E. degree from North Maharashtra University, Jalgaon, India, in 1999, the M.E. degree from Swami Ramanand Teerth Marathwada University, Nanded, India, in 2000 and Ph.D. degree from the Indian Institute of Technology (IIT) Kharagpur, India, in 2006. Currently, he is a Professor at DA-IICT, Gandhinagar, India. His research interests include speech processing, speech and speaker recognition, pattern recognition, wavelet signal processing, and infant cry analysis.

Dr. Patil has coedited a book with Dr. Amy Neustein (Editor-in-Chief, IJST, Springer-Verlag) on Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism, Springer, New York, USA. He is a PI/Co-PI for three DeitY and two DST sponsored projects.

Dr. Patil is an affiliate member of IEEE SLTC and member of IEEE, IEEE Signal Processing Society, IEEE Circuits and Systems Society (Awards), and ISCA.