# Combining Evidences from Mel Cepstral, Cochlear Filter Cepstral and Instantaneous Frequency Features for Detection of Natural *vs.* Spoofed Speech

*Tanvina B. Patel and Hemant A. Patil*

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India

{tanvina_bhupendrabhai_patel, hemant_patil}@daiict.ac.in

## Abstract

Speech synthesis and voice conversion techniques can pose threats to current speaker verification (SV) systems. For this purpose, it is essential to develop front end systems that are able to distinguish human speech *vs.* spoofed speech (synthesized or voice converted). In this paper, for the *ASVspoof 2015* challenge, we propose a detector based on combination of cochlear filter cepstral coefficients (CFCC) and change in instantaneous frequency (IF), (i.e., CFCCIF) to detect natural *vs.* spoofed speech. The CFCCIF features were extracted at frame-level and Gaussian mixture model (GMM)-based classification system was used. On the development set, the proposed features (i.e., CFCCIF) after fusion with Mel frequency cepstral coefficients (MFCC) features achieved an EER of *1.52* %, which is a significant reduction from MFCC (*3.26* %) and CFCCIF (*2.29* %) alone using *12-D* static features. The EER further decreases to *0.89* % and *0.83* % for delta and delta-delta features, respectively. Experimental results on evaluation set show that fusion of MFCC and CFCCIF works relatively well with an EER of *0.41* % for known attacks and *2.013* % EER for unknown attacks. On an average, fusion of MFCC and CFCCIF features provided relatively *best* EER of *1.211* % for the challenge.

**Index Terms**: CFCC, instantaneous frequency, spoofed speech, GMM, EER.

## 1. Introduction

An Automatic Speaker Verification (ASV) system accepts or rejects a claimed speaker's identity. In ideal cases, the speaker verification (SV) system should accept the claim for a true (i.e., genuine) speaker and reject the claim for an impostor. However, security and reliability of SV systems can be threatened by various spoofing attacks. The attacks can be due to impersonation, mimicking, replay, speech synthesis and voice conversion. Impersonation refers to human attacks caused by altering their voices (human mimicking) [1], [2]. Replay spoof is caused by reusing pre-recorded speech of the target/genuine speaker [3]. Spoofing due to speech synthesis uses text-to-speech (TTS) synthesis systems (generally Hidden Markov Model (HMM)-based TTS systems (HTS) and adapted HMM-based systems [4]- [5]) to produce natural and intelligible speech for a genuine speaker for any given text. Lastly, voice transformed or converted type of spoof is based on modifying a given speech of a source speaker to make it sound-like target (i.e., genuine) speaker [6], [7]. A detailed literature for various spoofing attacks can be found in [8].

Speaker adaptive speech synthesis and voice conversion attacks (unlike impersonation and replay attacks) can make use of easily accessible technology to produce a good quality spoofed speech for any target speaker. Thus, it is necessary to detect natural *vs.* spoofed speech (i.e., synthetic and voice converted). In addition, the detector must generalize to detect spoofed speech for any given attack. Impostor by speech synthesis was reported in the context of known spoof by HMM-based speech systems in [9]. A very recent work in this area is based on relative phase shift (RPS) that demonstrates reliable detection of synthetic speech and shows how RPS can be used to improve the security of SV systems [10]. In [11], effect of voice conversion spoofing techniques on the acceptance rates was studied, followed by anti-spoofing attack measures for SV systems [12]. Other studies on improvement of speech synthesis and voice conversion techniques confirm the exposure of the SV systems to spoofing threats.

For known attacks, in addition to magnitude-based features, research has initiated to use phase-based features to detect natural *vs.* spoofed speech. In [13], modified group delay phase features are used to detect voice converted speech. In [14], temporal modulation features are used for detecting synthetic speech. Earlier in [15], an auditory-based distortion measure was used to find the perceptual dissimilarity between speech segments and improve quality of synthetic speech by selecting speech sound units based on the auditory distortion measures. In this paper, we extend the use of cochlear filter cepstral coefficients (CFCC) based on wavelet transform-like auditory transform (AT) [16] and the related mechanisms that occurs in the cochlea of the human ear [17]. It is known that the envelope of each output of the cochlear filter, its instantaneous frequency (IF) and phase are important features used by auditory levels for speech perception (Chapter 8, pp. 403 [18]). Therefore, we propose CFCC plus IF (i.e., CFCCIF) features at the output of each subband filters to detect human and spoofed speech. The idea is that the human speech production system does not produce speech in a frame-by-frame pattern (rather in continuum) while feature extraction in speech synthesis and voice conversion is generally at frame-level. Thus, we propose capturing the feature variations across frames to detect natural *vs.* spoofed speech.

## 2. Proposed CFCCIF features

### 2.1. Cochlear filter cepstral coefficients (CFCC)

The parameter extraction procedure for auditory-based cepstral coefficients, consists of cochlear filterbank based on auditory transform (AT), hair cell function, nonlinearity and discrete cosine transform (DCT) [17]. The following sub-section describes in brief the AT and procedure for estimating the CFCC and proposed CFCCIF features.

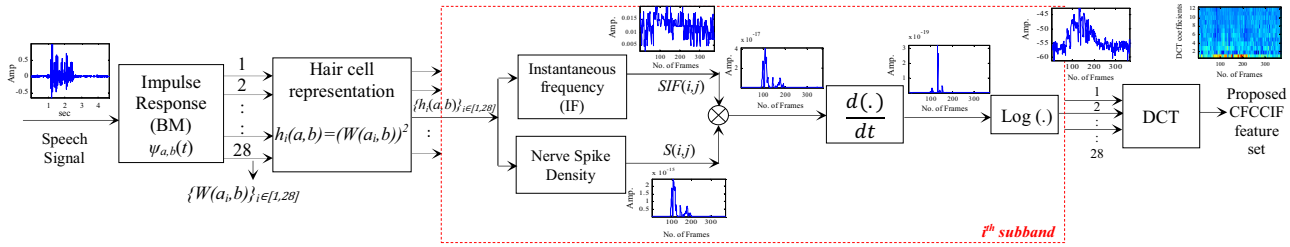



Figure 1: *Block diagram for proposed CFCCIF feature extraction scheme.*

### *2.1.1. Auditory Transform (AT)*

The AT was proposed in [16]. Let *s(t)* be the speech signal and the cochlear filter be $\psi(t)$. The AT of $s(t)$ (i.e., $W(a,b)$), *w.r.t.* $\psi(t)$ as impulse response of basilar membrane (BM) in the cochlea is defined as [16] - [17],

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

where
$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right). \tag{2}$$

In eq. (1), * is 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 to change the centre frequency while factor $b$ is the time shift or translation parameter. The energy remains equal for all $a$ and $b$. Hence, we have

$$\int_{-\infty}^{\infty} |\psi_{a,b}(t)|^2 = \int_{-\infty}^{\infty} |\psi(t)|^2 dt \,. \tag{3}$$

The cochlear filter is defined as [17],

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\left(\frac{t-b}{a}\right)^{\alpha} \exp\left[-2\pi f_L \beta\left(\frac{t-b}{a}\right)\right] \times \cos\left[2\pi f_L\left(\frac{t-b}{a}\right)+\theta\right]u(t-b). \tag{4}$$

Parameters $\alpha$ and $\beta$ determine the *shape* and *width* of cochlear filter and $\theta$ is selected such that the following *admissibility* condition for mother wavelet (i.e., $\psi(t)$), is satisfied [19]:

$$\int_{-\infty}^{\infty} \psi(t)dt = 0 \Rightarrow \psi(\omega)\big|_{\omega=0} = 0. \tag{5}$$

In particular, ∃ a number $C_\psi$ such that, $C_\psi = \int_0^{\infty} \frac{|\psi(\omega)|}{\omega} d\omega < \infty$ .

This means that the 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.,

$$a = \frac{f_L}{f_c}. \tag{6}$$

For the $i^{th}$ subband filter, its value of $a$ corresponding to $\{a_i\}$ needs to be pre-calculated for the required central frequency of the cochlear subband filters at band number $i\in[1,28]$.

### *2.1.2. 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 function of the hair cell describes this motion, i.e.,

$$h(a,b) = (W(a,b))^2; \qquad \forall W(a,b) \tag{7}$$

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,

$$S(i,j) = \frac{1}{d}\sum_{b=l}^{l+d-1} h(i,b), \quad l = 1,L,2L,....; \ \forall i,j \,, \tag{8}$$

where $d$ is the window length and $L$ is the window shift duration. The output of the above is further applied for scales of loudness functions as *cubic root* nonlinearity. However, use of CFCC in [17] suggests that the logarithmic nonlinearity operation is also appropriate. Finally, the discrete cosine transform (DCT) is applied to decorrelate the features.

## 2.2. Instantaneous frequency (IF) estimation

The IF of a signal $s(t)$ is defined as the derivative of the unwrapped phase of the analytic signal derived from $s(t)$. For a real signal $s(t)$, its complex analytic representation is given by,

$$s_a(t) = s(t) + js_h(t) \,, \tag{9}$$

where $s_h(t)$ is the Hilbert transform of the signal $s(t)$, given by the inverse Fourier transform (IFT) of $S_h(\omega)$, where,

$$S_h(\omega) = \begin{cases} +jS(\omega) & \omega < 0 \\ -jS(\omega) & \omega > 0 \end{cases}. \tag{10}$$

Thus, the amplitude (Hilbert) envelope of $s_a(t)$ is given by,

$$|s_a(t)| = \sqrt{s^2(t) + s_h^2(t)} \,, \tag{11}$$

instantaneous phase is $\phi(t) = \tan^{-1}\left(s_h(t) \big/ s(t)\right)$, and IF derived from derivative of *unwrapped* instantaneous phase, is given as,

$$IF = \frac{d}{dt}\left(\phi(t)\right). \tag{12}$$

## 2.3. Estimation of CFCCIF features

The block diagram of the proposed CFCCIF features is shown in Figure 1. Similar to nerve spike density estimation, for $d$ window length and $L$ window shift, the IF is obtained as,

$$SIF(i,j) = \frac{1}{d}\sum_{b=l}^{l+d-1} IF(h(i,b)), \quad l = 1,L,2L,....; \ \forall i,j \tag{13}$$

To use both envelope structure and IF information, the framewise IF features (eq. 13) are multiplied with the corresponding nerve spike density envelope (eq. 8). Thus, IF obtained in silence regions will be suppressed. To capture the transient information, the *change* in envelope and IF between consecutive frames is estimated through derivative operation followed by logarithm. This is repeated for all subbands, i.e., $i\in[1,28]$ (shown by dotted region in Figure 1). Finally, DCT is applied framewise to get CFCCIF features. Figure 2 shows a speech signal (natural speech), the energy at outputs of the cochlear filterbanks (CFCC) and energy after embedding the information in the IF of the speech signal (CFCCIF). It is seen that adding IF information and using the change across frames enhances the information in the representation of the CFCC (as shown by dotted regions in Figure 2(b) –2(c)).

Table 2. *The score-level fusion % EER obtained on development dataset for D1, D2 and D3-dimensional feature vector.*

| Features with score-level fusion | Dimension (*D*) of feature vector | EER (%) for varying values of $\alpha_f$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
| MFCC+CFCC | D1: *12*-static | 3.26 | 2.86 | 2.66 | 2.52 | **2.43** | 2.57 | 2.72 | 3.03 | 3.55 | 3.97 | 4.55 |
| MFCC+(CFCCIF) | | 3.26 | 2.72 | 2.40 | 2.03 | 1.77 | 1.60 | **1.52** | 1.57 | 1.72 | 1.92 | 2.29 |
| MFCC+CFCC | D2: *12*-static +*12* delta | 2.17 | 1.83 | 1.54 | 1.40 | **1.32** | 1.32 | 1.46 | 1.63 | 1.89 | 2.23 | 2.60 |
| MFCC+(CFCCIF) | | 2.17 | 1.83 | 1.46 | 1.23 | 1.03 | 0.97 | **0.89** | 0.89 | 0.97 | 1.14 | 1.40 |
| MFCC+CFCC | D3: *12*-static +*12* delta + *12* (delta-delta) | 1.60 | 1.32 | 1.14 | 0.97 | **0.89** | 0.89 | 0.92 | 1.00 | 1.17 | 1.34 | 1.54 |
| MFCC+(CFCCIF) | | 1.60 | 1.37 | 1.14 | 1.00 | 0.86 | **0.83** | **0.83** | 0.92 | 1.03 | 1.17 | 1.52 |

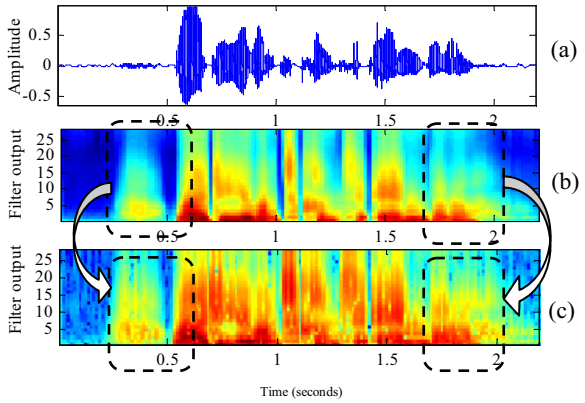



Figure 2: *(a) A natural utterance (16 kHz) provided from the challenge* [20]*, (b) CFCC: the output of 28 cochlear subband filters and (c) CFCCIF: the output of 28 cochlear subband filters with the IF information.*

## 3. Experimental results

For the *ASVspoof 2015* challenge [20], a classifier/detector is built to deal with given spoofing attacks. The database was provided as a part of the challenge by the organizers. Brief details of the database are given in Table 1. Details of the spoofing algorithms (S) are provided in [21]. 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 attacks. The S3, S4 and S10 are based on speech synthesis and remaining on voice conversion system. The state-of-the-art Mel frequency cepstral coefficients (MFCC), CFCC and CFCCIF features are extracted on three different dimensions of feature vector, i.e., **D1**: *12-D* static features, **D2**: *24-D* (12 static+*12* delta), **D3**: *36-D* (12 static+*12*delta+*12* (delta-delta)).

Table 1. *Statistics of the dataset provided for the ASVspoof 2015 challenge* [22].

| | No. of speakers | | No. of utterances | |
|---|---|---|---|---|
| Dataset | Male | Female | Genuine | Spoofed |
| Training | 10 | 15 | 3750 | 12625 |
| Development | 15 | 20 | 3497 | 49875 |
| Evaluation | 20 | 26 | 193404 | |

### 3.1. Model training and score-level fusion

In this paper, we use Gaussian Mixture Model (GMM) with *128* mixtures for modeling the classes corresponding to natural and spoofed speech. GMM for natural speech is built using entire training dataset of *3750* genuine (i.e., natural) utterances. Similarly, GMM for spoofed speech is built with *12625* spoofed training utterances. Final scores are represented in terms of log-likelihood ratio (LLR). The decision of the test speech being human or spoofed is based on the LLR, i.e.,

$$LLR = \log(LLk\_Model1) - \log(LLk\_Model2), \quad (14)$$

where *LLk_Model1* and *LLk_Model2* are the likelihood scores from the GMM for the human speech and spoofed speech, respectively. In our study, we have extracted features from *25 ms* of frame with a shift of *50 %* and using *28* subband filters for MFCC, CFCC and proposed CFCCIF features. To utilize possible complementary information in MFCC and CFCCIF features (i.e., the spectral information in MFCC and IF information in CFCCIF), we use their score-level fusion, i.e.,

$$LLk_{combine} = (1-\alpha_f)LLk_{MFCC} + \alpha_f LLk_{feature2} \quad (15)$$

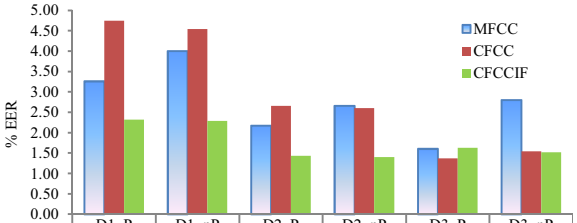
where $LLk_{MFCC}$ and $LLk_{feature2}$ is the log-likelihood score of MFCC and CFCC or CFCCIF, respectively. The weights of the scores are decided by the fusion parameter $\alpha_f$.

### 3.2. Performance measure

Detection Error Tradeoff (DET) curve is used to measure the performance of MFCC, CFCC and CFCCIF features [23]. It gives uniform treatment to both false acceptance and miss rejection rate for evaluation of system performance. In DET curve, the operating point where false acceptance rate and miss rejection rate becomes equal is referred to as Equal Error Rate (EER). The false acceptance rate and miss rejection rate at threshold is calculated as per evaluation plan of *ASVspoof 2015* challenge [22]. The organizers of the challenge have used Bosaris toolkit to compute % EER [24].

### 3.3. Effect of pre-emphasis

To study feature dependence due to pre-emphasis ($a_{pre}$=*0.97*) on the speech signal, the % EER was obtained for individual systems using MFCC, CFCC and CFCCIF features for *D1*, *D2* and *D3* set of features. As shown in Figure 3, MFCC features have sensitive dependence to pre-emphasis (*P*), i.e., for no pre-emphasis (*nP*), the % EER increases significantly for all sets of feature dimensions. On the other hand, the % EER of CFCC and CFCCIF (with *P* or *nP*) is almost constant on all feature




| | 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 |

Figure 3: *Effect of pre-emphasis on % EER, using MFCC, CFCC and CFCCIF features (P=pre-emphasis and nP=no pre-emphasis on speech signal).*

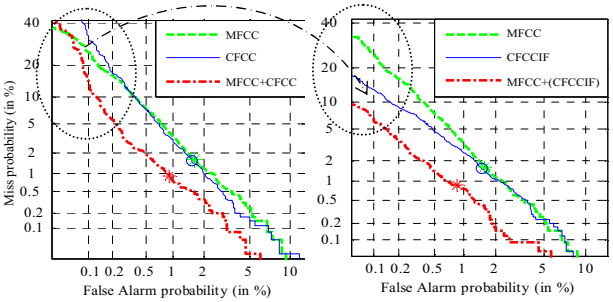
Table 3: *% EER results of primary submission (i.e., for 36-D MFCC+CFCCIF ($\alpha_f$ =0.6)) of ASVspoof 2015 challenge* [21].

| Submission | Known attacks (% EER) | | | | | Unknown attacks (% EER) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | Avg. |
| **A**: DA-IICT | 0.1013 | 0.8629 | 0.0000 | 0.0000 | 1.0753 | 0.8462 | 0.2416 | 0.1417 | 0.3463 | 8.4900 | **1.211** |
| Average (Proposed) | 0.407899 | | | | | **2.013162** | | | | | |
| Avg. of *16* submissions | 3.337 | | | | | 9.294 | | | | | 6.316 |

dimensions. In fact, on an average, the CFCC and CFCCIF features perform better without pre-emphasis on speech signal. Thus, the advantage of CFCC/CFCCIF is that *no* pre-emphasis is needed due to embedded bandpass filtering (i.e., due to the admissibility condition of cochlear filter function $\psi(t)$, eq. (5)).

### 3.4. Results and discussions on the development set

Results for MFCC, CFCC and proposed CFCCIF features are shown in Table 2. The cochlear filter shape parameters are optimized to $\alpha$=*3* and $\beta$=*0.035* through intensive experiments. These values of $\alpha$ and $\beta$ gives a narrow shape to the cochlear filter which may help to capture speaker-specific information. Assuming that synthesized/voice converted speech does not exactly match to target human speaker, the speaker differences will exist. Table 2 shows that the CFCCIF features produce much lower % EER than MFCC and CFCC which confirms that the differences in spoofed speech and natural speech are captured, i.e., the CFCCIF features are different for synthetic and voice converted speech than the human speech. CFCCIF features have both cepstral and phase information and hence, they perform better than MFCC features used alone.




(a) (b)

Figure 4: *(a) DET curve for MFCC (--green), CFCC (blue) and their score-level fusion with $\alpha_f$=0.4 (-.-red), (b) DET curve for MFCC (--green), CFCCIF (blue) and their score-level fusion with $\alpha_f$ =0.6 (-.-red).*

Furthermore, the score-level fusion of these features was done as per eq. (15). It was observed that for almost equal weighted fusion of MFCC and CFCCIF scores, the % EER of MFCC (*3.26*) and CFCCIF (*2.29*) reduces to *1.52* for *D1* set of features. A similar trend is observed for *D2* and *D3* features. It is observed that the weight of fusion (i.e., $\alpha_f$) decides the contribution of individual system (e.g., $\alpha_f$ = *0.4* for CFCC and $\alpha_f$=*0.6* for CFCCIF). Therefore, it can be said that the CFCCIF features have more contribution in decreasing the % EER. Table 2 shows that the proposed method captured the *complementary* information that was not evident from MFCC alone. This is also evident from the oval dotted regions shown in DET curves as in Figure 4 (a)-(b).The score-level fusion is denoted by '+' in Figure 4. The % EER, i.e., when the miss probability (human speech is detected as spoofed) and false alarm probability (spoofed speech accepted as genuine speech) is equal, is very less for fusion of the MFCC and CFCC features as in Figure 4 (a) and for MFCC and CFCCIF features as in Figure 4 (b). To obtain relatively least % EER on the development dataset, MFCC features are obtained on pre-emphasized speech and CFCC and CFCCIF features were obtained without pre-emphasis (as CFCC inherently employs bandpass filter $\psi(t)$ eq. (5)). It was observed that with this combination, % EER is ***0.83*** as in Table 2. On the other hand, with pre-emphasis on both features, the % EER was *0.86*. The difference is rather small on the development set. However, the difference would be significant on large test datasets.

### 3.5. Results on evaluation data

The primary system submitted by our team is denoted by "A" [21]. Table 3 indicates the results in % EER of the *ASVspoof* evaluation data for known and unknown attacks. The spoofing detection scores were based on fusion of *36-D* features (*12* static+*12* delta+*12*delta-delta) MFCC and CFCCIF with factor *0.6 for CFCCIF features*. The proposed CFCCIF feature gave *0.41* % EER for known attacks. On the other hand, the EER for unknown attacks was *2.013 %* which was the least among all the *16* submissions at the challenge [21]. It was observed that speech synthesis attacks (S3, S4) were easily identified by the detector for known attacks. On the other hand, speech synthesis by unknown spoof, i.e., by MARY Text-To-Speech (MaryTTS) system [25] was most difficult to detect. The EER for voice conversion spoof was almost less than *1 %* in all known and unknown cases. In the challenge, the best EER for known attacks was that of system *D*, i.e., *0.003* % which increased to *5.231* % for unknown attacks [21]. Thus, the countermeasures of the system *D* might be biased towards the prior information used while training [21]. This shows that the proposed CFCCIF feature was robust to unknown attacks and almost independent of the nature and type of attacks.

## 4. Summary and conclusions

The paper shows the improvement on combining MFCC, CFCC and CFCCIF to detect natural *vs.* spoofed speech. In addition to cepstral features, the use of IF to capture perceptual information proves to be very effective. It has been observed that on the standard dataset provided for the challenge, the score-level fusion of the MFCC and CFCCIF features gave quite low % EER for known attacks and relatively best lowest % EER for unknown attacks among the various submissions at the *ASVspoof* 2015 challenge, which makes the proposed countermeasure suitable for real case scenario of spoofing attacks. It has been shown that CFCC performs almost similarly to MFCC for speaker identification problem in clean conditions [17]. However, CFCC features outperform MFCC features under noisy or signal degradation conditions. Therefore, the authors would like to explore use of CFCCIF features for *robustness* in presence of additive or channel noise and its relative effects of various types of spoofed speech.

## 5. Acknowledgements

The authors thank Dept. of Electronics and Information Technology (DeitY), Govt. of India, for sponsored project, "Development of Text-to-Speech (TTS) System in Indian Languages (Phase-II)" and the authorities of DA-IICT Gandhinagar.

# 6. References


[1] P. Perrot and G. Chollet, "Helping the forensic research institute of the French Gendarmerie to identify a suspect in the presence of voice surgery," in *Forensic Speaker Recognition*, A. Neustein and H. A. Patil, (Eds.), Springer, pp. 469–503, 2012.

[2] Y. W. Lau, M. Wagner and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proc. Int. Symp. on Intell. Multimedia, Video, Speech Process.*, Hong kong, pp. 145-148, 2004.

[3] F. Alegre, A. Janicki and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *Proc. Int. Conf. of the Biometrics Special Interest Group (BIOSIG '14)*, Darmstadt, Germany, pp. 157-168, 2014.

[4] K. Tokuda, H. Zen and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Workshop on Speech Synthesis (SSW '02)*, pp. 227-230, 2002.

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

[6] Y. Stylianou, "Voice transformation: A survey," in *Proc. Int. Conf. on Acous., Speech and Sig. Process. (ICASSP '09)*, Taipei, Taiwan, pp. 3585-3588, 2009.

[7] J. -F. Bonastre, D. Matrouf and C. Fredouille, "Transfer function-based voice transformation for speaker recognition," in *Proc. IEEE Speaker Lang. Recogn. Workshop (Odyssey '06)*, Toledo, pp. 1-6, 2006.

[8] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Comm.,* vol. 66, p. 130–153, February 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. European Conf. on Speech Comm. and Technol. (EUROSPEECH '99)*, Budapest, Hungary, pp. 1223-1226, 1999.

[10] 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 Trans. on Audio, Speech, and Lang. Process.,* vol. 20, no. 8, pp. 2280-2290, Oct 2012.

[11] J. F. Bonastre, D. Matrouf and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH '07)*, Antwerp, Belgium, pp. 2053-2056, 2007.

[12] Z. Wu and H. Li, "Voice conversion and spoofing attack on speaker verification systems," in *Proc. Asia Pacific Signal and Inf. Process. Assoc. Annual Summit and Conf. (APSIPA ASC '13)*, Taiwan, pp. 1-9, 2013.

[13] Z. Wu, E. S. Chng and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH '12)*, Portland, Oregon, USA, pp. 1700-1703, 2012.

[14] 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 '13)*, Vancouver, BC, Canada, pp. 7234 - 7238, 2013.

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

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

[17] 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.

[18] T. F. Quatieri, Discrete-Time Speech Signal Processing, NJ: Prentice Hall, Engewood Cliffs, 2002.

[19] S. G. Mallat, A Wavelet Tour of Signal Processing, Academic Press, 1998.

[20] ASVspoof 2015: Automatic Speaker Verification Spoofing and Countermeasures Challenge, [Available Online]: http://www.spoofingchallenge.org/ {Last accessed *10th* June. 2015}.

[21] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, Md. Sahidullah and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge", in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH '15)*, Dresden, Germany, 2015 (To appear).

[22] Z. Wu, T. Kinnunen, N. Evans and J. Yamagishi, "ASVspoof 2015: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan", [Available Online]: http://www.spoofingchallenge.org/asvSpoof.pdf, pp. 1-5, December 19, 2014.

[23] A. Martin, G. Doddington, T. Kamm and M. Ordowski, "The DET curve in assessment of detection task performance," in *Proc. European Conf. on Speech Comm. and Technol. (EUROSPEECH '97)*, Rhodes, Greece, pp. 1895-1898, 1997.

[24] BOSARIS Toolkit, [Available Online].: https://sites.google.com/site/bosaristoolkit/ {Last accessed *20th* Feb. 2015}.

[25] The MARY Text-to-Speech System (MaryTTS), [Available Online]: http://mary.dfki.de/ {Last accessed *19th* March 2015}.