Novel Energy Separation Based Frequency Modulation Features For Spoofed Speech Classification

Madhu R. Kamble and Hemant A. Patil
Speech Research Lab
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT)
Gandhinagar, India
{madhu_kamble, hemant_patil}@daiict.ac.in

***Abstract*—Speech Synthesis (SS) and Voice Conversion (VC) methods provides a great risk for Automatic Speaker Verification (ASV) system. In this paper, we tried to find the difference between natural and spoofed speech signals using Teager Energy Operator-based Energy Separation Algorithm (TEO-ESA). Here, we exploit the contribution of Amplitude Envelope (AE) and Instantaneous Frequency (IF) in each narrowband filtered signals energy via ESA to capture possible changes in a *temporal* and *spectral* envelope of the synthetic speech signal generated by the machines as opposed to natural signals. Furthermore, IF was used for classification of natural *vs.* spoof speech with Gaussian Mixture Model (GMM) as a classifier. These findings may assist to distinguish these two speeches and provide an aid to alleviate possible impostor attacks in voice biometrics. The experiments are done on ASV Spoof 2015 Challenge database. We have compared proposed Energy Separation Algorithm-Instantaneous Frequency Cosine Coefficients (ESA-IFCC) with Mel Frequency Cepstral Coefficients (MFCC) features. On the development set, MFCC alone gave an Equal Error Rate (EER) of (6.98 %) and ESA-IFCC gave (5.43 %) with 13-D static features. With score-level fusion of MFCC and ESA-IFCC EER reduced to 3.45 % on static feature vector. The EER decreases further to 2.01 % and 1.89 % for Δ and $\Delta\Delta$ features. On evaluation set, the overall average error rate for known and unknown attacks was 6.79 % for ESA-IFCC and was significantly better than the MFCC (9.15 %) and their score-level fused EER (7.16 %).**

***Index Terms*—Automatic Speaker Verification, Spoofing Attacks, Teager Energy Operator, Amplitude Envelope, Instantaneous Frequency.**

I. Introduction

Spoofing attacks replicate a person's identity from his or her voice to get access to a sensitive or protected system [1]. Recent developments in speech technology related to SS and VC present a great threat to speech-based biometric system [1]. The various spoofing attacks include, speech synthesis (SS) [2], voice conversion (VC) [3], replay [4], impersonation [5]. To playback pre-recorded speech samples to get the access of a system is a replay attack [6]–[8]. An impersonation tries to mimic a genuine target speaker [5], [9]. The other machine-generated techniques include Text-to-Speech (TTS) synthesis [10] and voice conversion [11]–[13].

978-1-5386-2241-4/17/$31.00 ©2017 IEEE

The ASV spoof 2015 challenge at INTERSPEECH 2015 brought common database that has SS and VC speech signals. The general countermeasure approach was to detect genuine speech from other spoofed speech was attempted in this ASV spoof 2015 challenge [14]. Various feature extraction methods, such as, magnitude-based, phase-based, the combination of amplitude and magnitude and prosodic feature were reported to detect SS and VC spoofing attacks. The Cochlear Filter Cepstral Coefficients-Instantaneous Frequency (CFCCIF) [15], and Constant Q Cepstral Coefficients (CQCC) [16] are considered as the state-of-the-art features on ASV spoof 2015 challenge database. These features performed as best detection, especially for unknown attacks in the evaluation set. This might suggest that the feature extraction has much more important role to design of spoofing countermeasures rather than the advanced or complex classifiers [17]. In this paper, we propose to analyze natural *vs.* spoofed speech by exploiting Teager Energy Operator-based Energy Separation Algorithm (TEO-ESA) [18], [19]. ESA uses nonlinear energy operator to track instantaneous energy of source generating amplitude and frequency modulation (AM-FM) signal and separate it into its amplitude and frequency components [20]. AM conveys both phonemic and speaker-dependent information, whereas FM is perceptually important [21]. Here, we exploit the contribution of Amplitude Envelope (AE) and Instantaneous Frequency (IF) in each subband filtered signals energy via ESA. Both AE and IF are computed from the narrowband components of speech signal. However, spoofed speech are not generated in the same way as natural speech and hence, we can observe the perceptual difference (to a certain extent) that occur in both type of speech (i.e., natural *vs.* spoof). The mean and standard deviation are used to calculate the difference between natural and spoofed speech signal. It is observed that standard deviation of IF with large number of subband filtered signals shows much more difference that makes it as a distinguishable feature to classify natural *vs.* spoofed speech.

We have used IF to extract the feature and discarded AE. As IF shows the difference between natural and spoofed speech signal. Furthermore, the IF of each narrowband filtered signal is frame-blocked and averaged it followed by Discrete Cosine

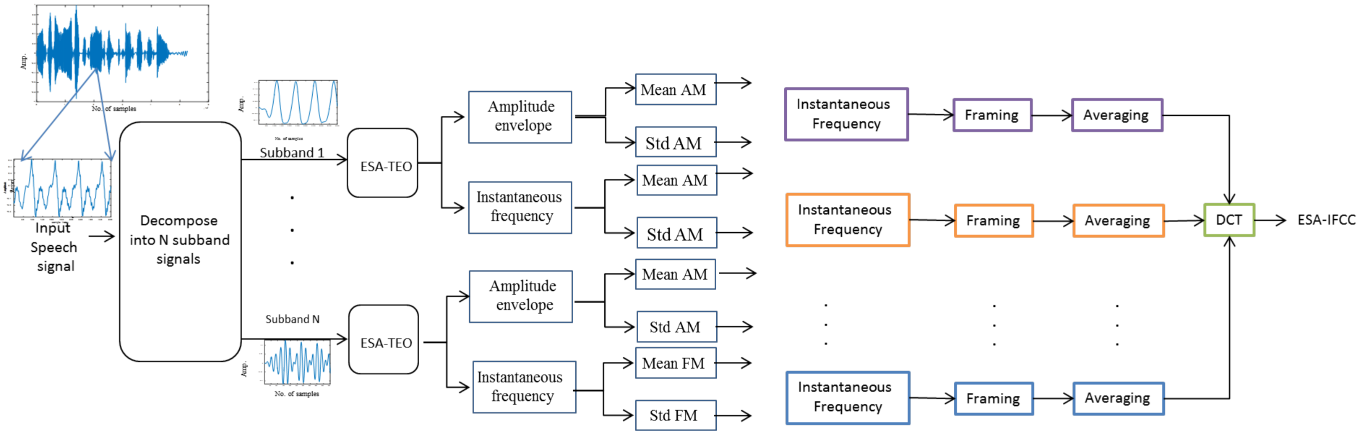

Fig. 1. Schematic diagram for extraction of proposed mean and standard deviation by ESA.

Transform (DCT) and extracted the proposed Energy Separation Algorithm-Instantaneous Frequency Cosine Coefficients (ESA-IFCC) feature set.

II. Background

According to the Teager, modeling the time-varying vortex-flow is a challenging task [22]. A simple algorithm is derived by Teager, that uses a nonlinear energy tracking operator called as Teager Energy Operator (TEO) in discrete-time domain for speech signal analysis both in continuous-time and discrete-time domain [23].

A. Teager-Kaiser Energy Operator

The Hooke's law of elasticity and second law of motion by Newton with mass *m* and spring constant *k* for an oscillator, governs a system dynamics that is denoted by a second-order differential equation as:

$$\frac{d^2x}{dt^2} + \frac{k}{m}x = 0, \tag{1}$$

The solution for Eq. 1 is a sinusoidal signal, with $x(t) = Acos(\Omega t + \phi)$, where A, Ω and ϕ are the amplitude, frequency in rad/sec and initial phase in rad of oscillation. The total energy of a system, E, is the summation of potential and kinetic energy and is given by:

$$E = P.E + K.E, \tag{2}$$

$$E = \frac{1}{2}kx^2 + \frac{1}{2}m\dot{x}^2, \tag{3}$$

$$E = \frac{1}{2}m\Omega^2 A^2, \tag{4}$$

where $\Omega = d\phi(t)/dt$. Considering this analysis, Teager and Kaiser, proposed the *Teager-Kaiser Energy Operator* for discrete-time signal [23], i.e.,

$$\Psi_d\{x(n)\} = x^2(n) - x(n-1)x(n+1), \tag{5}$$

$$E_n \approx A^2\omega^2, \tag{6}$$

where E_n provides running estimate of the signal's energy.

B. Energy Separation Algorithm (ESA)

The contribution of amplitude $a[n]$ and frequency $\omega[n]$ of signal were estimated using TEO framework [24], [25]. The *Energy Separation Algorithm (ESA)* was developed that tracks instantaneous energy of the source generating AM-FM signal with nonlinear energy operator to separate the signal into amplitude and frequency components. The ESA provides amplitude envelope (AE) and instantaneous frequency (IF) of a speech signal. Energy of the speech signal is a function of amplitude and frequency components [26]. As the speech signal is multicomponent, i.e., containing several resonances, ESA is applied to a single speech resonance. The IF $\omega[n]$ and AE $a[n]$ of the AM-FM modulated signal $x[n]$ for i^{th} subband at any time instant is given by [27]:

$$a_i[n] \approx \frac{2\Psi_d\{x[n]\}}{\sqrt{\Psi_d\{x[n+1]) - x[n-1]\}}}, \tag{7}$$

$$\omega_i[n] \approx arcsin\sqrt{\frac{\Psi_d\{x[n+1] - x[n-1]\}}{4\Psi_d\{x[n]\}}}. \tag{8}$$

Eq. (7) and Eq. (8) are w.r.t. symmetric approximation of derivative operation required in the continuous-time version of TEO, i.e.,

$$y[n] = \frac{x(n+1) - x(n-1)}{2}. \tag{9}$$

For transmitting information, modulations of amplitude and frequency are used extensively in communication systems. The AM and FM models the representation of time-varying amplitude and frequency patterns in speech resonances. The AE signals of different formants are highly correlated for voiced speech and have a specific structure. The multipulse excitation signals for AE of different formant bands are expected to be coupled with voiced speech and loosely coupled for unvoiced speech [28].

The IF of a signal is interpreted as the frequency of the sinusoid that *locally* fits the given frequency of the signal. The IF is modeled as the combination of the slow and fast-varying components. The average formant frequency are modeled

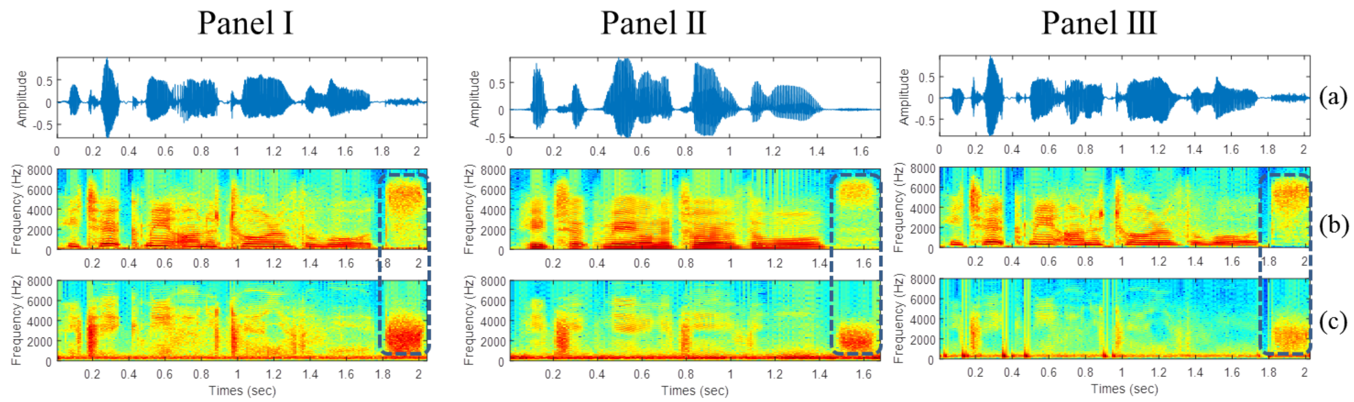

Fig. 2. Spectrographic analysis of a speech signal Panel I for natural, Panel II for SS and Panel III for VC speech. (a)time-domain speech signal (b) corresponding spectrogram of (a) and (c) the spectral energy density of 40 subband signals.

by slow-varying component, and the fast-varying component models frequency variations near the formant frequency. Next, we have computed mean (μ) and standard deviation (σ) over the instantaneous amplitude and frequency of AM-FM signal of N bandpass (subband) filtered signals. In particular, mean and standard deviation of subband filtered signal are given as:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{10}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}. \tag{11}$$

III. Analysis of Amplitude Envelope and Instantaneous Frequency

Fig. 1 shows the block diagram for extraction of mean and standard deviation by ESA. The input speech signal is passed through a bank of filters to obtain N bandpass filtered signals. The filterbank we have used here is Butterworth filter. The ESA-TEO is applied on each N number of bandpass filtered signals to obtain corresponding AEs and IFs of the narrowband signal. To compute the importance of AEs and IFs, we have calculated their mean and standard deviations as per Eq.(10) and Eq. (11) of each utterance for natural, and spoofed speech signals.

TABLE I
Mean and stabdard deviation of AM and FM for training set of ASV spoof 2015 database

	40 Butterworth filters			
Speech	meanAM	stdAM	meanFM	stdFM
Natural	0.0012	0.0027	2027	28
SS1	0.0011	0.0023	2026	30.31
SS2	0.0011	0.0023	2026	29.94
VC1	0.0014	0.0023	2026	25.17
VC2	0.00094	0.0022	2026	23.23
VC3	0.00088	0.0019	2026	23.71

We performed an experiment based on mean and standard deviation on the development set of the ASV spoof 2015 challenge database as discussed in Section IV-A. It can be observed from the Table I, that for SS (i.e., SS1 and SS2 speech synthesis spoofs based on hidden Markov model (HMM)-based methods) the standard deviations of IF are almost the same for 40 number of subband filtered signals. This observation indeed shows that for extracting IF of a particular frequency, we require more number of subband filtered signals as speech is multicomponent signals. For VC-based spoofed speech (i.e., VC1, VC2 and VC3) shows relatively significant difference with natural speech signals with standard deviations of IF. From Table I, it is observed that VC1 that is frame selection-based VC method (this VC method is considered as mirror of natural speech that makes it difficult to classify) are more biased towards natural speech that is not in the case of VC2 (spectral slope shifting) and VC3 (Festvox).

This information of instantaneous frequency made us to further analyze the IF and AE and to propose a new feature from the IF that could make classification of natural and spoofed speech signals. From Table I, we conclude that IF plays important role for classification as compared to that of AE. To use IF as a feature, we have done few modification in the Fig. 1 so that we can extract features from TEO-based ESA. Fig. 1 shows the block diagram of proposed feature extraction with TEO-based ESA-IFCC feature set. With 40 channel linearly-scaled 3^{rd} order Butterworth filterbank ESA-IFCC features were extracted for the frequency range of minimum frequency 100 Hz and maximum frequency 7800 Hz. For each narrowband component, AM-FM components were computed using TEO-based ESA. The computation of IF was done for each narrowband component and furthermore, these IFs were averaged for short-time windows of 20 ms duration, with a window shift of 10 ms, to obtain 20-D feature vector. The DCT was applied on IFCs and retained first 20 coefficients in the transformed-domain to obtain ESA-IFCC feature set. To obtain high-dimensional feature set these 20-D ESA-IFCCs, were appended with their dynamic (delta) and acceleration (double-delta) features.

The Butterworth filter provides a response with maximally flat and used to bandpass filter the signal in the passband. The frequency scale of the Butterworth filterbank is kept to be with linear frequency scale. As linear scale almost have equal bandwidth over all the frequency range that helps to capture the higher formants present in the higher frequency range. The spectrographic analysis of natural, SS and VC speech signals is shown in Fig. 2 denoted by (panel I, panel II and panel III). The time-domain speech signal is shown in Fig. 2(a) and its corresponding spectrogram are shown in Fig. 2(b), whereas the spectral energy density obtained after 40 subband filtered signals is shown in Fig. 2(c). It can be observed that the lower frequency regions are more emphasized in natural, SS and VC speech signal.

IV. Experimental Results

A. Experimental Setup

The experiments are conducted on the statistically meaningful ASV spoof 2015 database [1]. The database have three subsets, i.e., training, development and evaluation set. Table II summarizes speaker and utterance information for each subset.

TABLE II
Number of speakers and utterances in different datasets

Subset	# Speakers		# Utterances	
	Male	Female	Genuine	Impostor
Training	10	15	3750	12625
Development	15	20	3497	49875
Evaluation	20	26	9404	193404

B. Feature Extraction Parametrization

We have compared the ESA-IFCC features with MFCC features. The features are extracted as per following details:

- **MFCC:** These feature sets were extracted with 28 subband filter and using 25 ms frame size with frame shift of 10 ms. Features were extracted with 39-D (static+delta+double-delta) feature vector.
- **ESA-IFCC:** This feature set is extracted with 40 subband filtered. Only first 20 static coefficients were extracted appended with delta and double-delta resulting in 60-D feature vector.

We have used Gaussian Mixture Model (GMM) for modeling the classes corresponding to genuine, and spoofed speech utterances of the training set. The final scores were represented in terms of Log-Likelihood Ratio (LLR). The decision of the test speech being genuine or spoofed is based on the scores of LLR:

$$LLR = log\frac{P(X|H_0)}{P(X|H_1)}, \tag{12}$$

where $P(X|H_0)$, and $P(X|H_1)$ are likelihood scores of genuine and spoof trials (with hypothesis H_0 and H_1), respectively.

To explore the possible complementary information captured by MFCC and ESA-IFCC features, we use their score-level fusion, i.e.,

$$LLK_{combine} = (1-\alpha)LLK_{feature1} + \alpha LLK_{feature2}, \tag{13}$$

The fusion parameter (α) lies between $0 < \alpha < 1$ to decide the weight of scores.

C. Results on Development Set

Results for MFCC and proposed ESA-IFCC feature set are shown in Table III with different number of mixtures in GMM. We have used 16, 32, 64 and 128 number of mixtures in GMM for computation of the models. It is observed that for proposed ESA-IFCC feature set with increased number of mixtures in GMM their is an decreases in EER. However, with MFCC features the EER increases ranging between 5.11 % to 7.47 % while, the EER of ESA-IFCC feature remains constant and ranging between 6.64 % to 6.73 %. From these findings ESA-IFCC features produce much lower % EER and are capable to classify genuine *vs.* spoofed speech than the MFCC alone.

TABLE III
Results on development set with different number of mixtures in GMM

GMM	MFCC	ESA-IFCC
16	5.11	6.64
32	7.10	6.79
64	7.70	6.72
128	7.47	6.73

TABLE IV
Results in EER (%) on development set of ASV spoof 2015 challenge database with Butterworth filterbank

Feature Set	static (S)	S+Δ	S+Δ+$\Delta\Delta$
MFCC	6.98	6.75	7.47
ESA-IFCC	5.43	6.22	6.73
MFCC+ESA-IFCC	**3.45**	**2.01**	**1.89**

Results of MFCC, proposed ESA-IFCC with their score-level fusion on the development set of ASV spoof 2015 challenge database are shown in Table IV. By using static feature vector, ESA-IFCC, achieved much lower EER (5.43 %) compared to the MFCC (6.98 %). However, for double-delta feature vector, the EER is not decreased. As the proposed feature are phase-related features for high-dimensional coefficients there may be more variations in the phases because of that it may not capture the changes. The MFCC is magnitude-based feature and ESA-IFCC is phased-based, score-level fusion captures the complementary information of both the features. When fused at a score-level, the EER is reduced to 3.45 % on static 2.01 % on static+Δ and 1.89 on static+Δ+$\Delta\Delta$. The DET curves of the same features (as in Table IV and Table V) are shown in Fig. 3. The Fig. 3(a) shows the individual DET curve on development set.

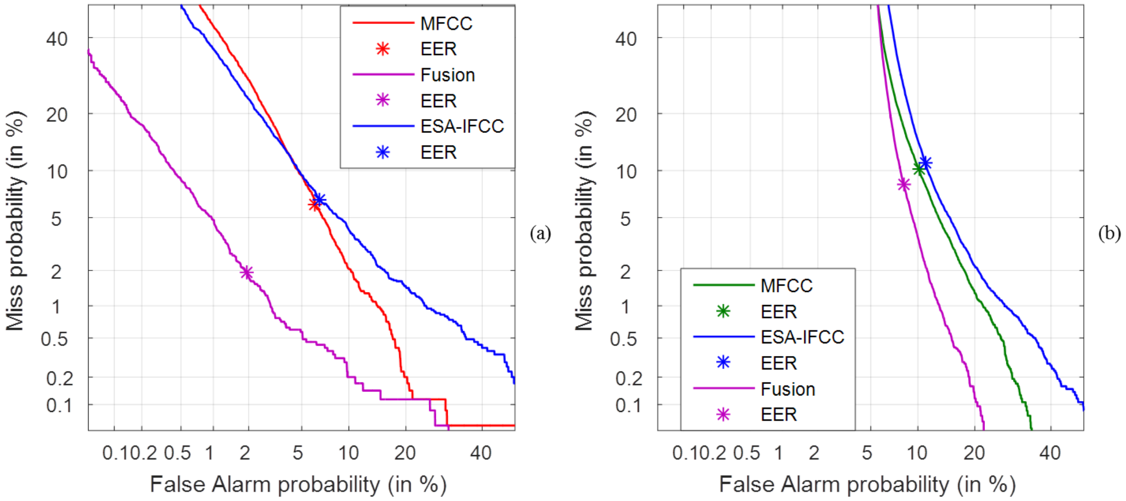

Fig. 3. The DET curves for MFCC, ESA-IFCC and their score-level fusion on (a) development set and (b) evaluation set.

TABLE V
RESULTS IN % EER ON EVALUATION SET FOR ASV SPOOF 2015 CHALLENGE DATABASE ON EACH SPOOFING ATTACK WITH BUTTERWORTH FILTER

Features	Known Attacks						Unknown Attacks						All Avg.
	S1	S2	S3	S4	S5	Avg.	S6	S7	S8	S9	S10	Avg.	
MFCC	2.34	9.57	**0.00**	**0.00**	9.01	4.18	7.73	4.42	0.3	5.17	52.99	14.12	9.15
ESA-IFCC	2.68	4.87	**0.00**	**0.00**	12.87	4.08	10.9	2.4	3.57	3.33	**37.37**	**9.514**	**6.79**
MFCC+ESA-IFCC	**0.78**	**3.39**	**0.00**	**0.00**	**5.45**	**1.92**	**4.19**	**1.22**	**0.11**	**1.80**	54.73	**12.41**	7.16

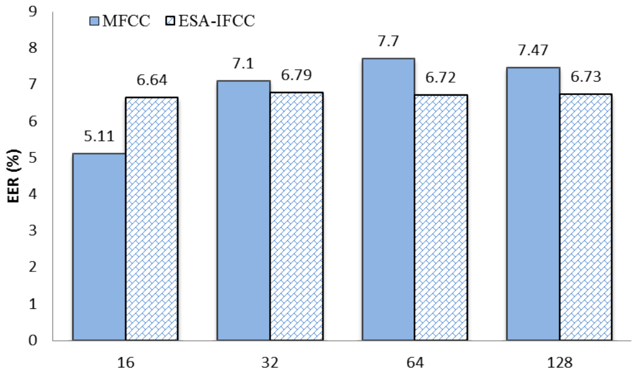

Fig. 4. Results with different number of mixtures in GMM. We have computed with 16, 32, 64 and 128 number of mixtures.

D. Results on Evaluation Dataset

the results on evaluation set of ASV spoof 2015 Challenge database are shown in Table V. The evaluation set consists of 10 spoofing algorithms among which 5 algorithms were same as used in development set and known as known attacks. The other 5 attacks were introduced directly to the challenge and were known as unknown attacks. It was observed that S3 and S4 that are SS-based attacks and are easy to detect for the case of known attacks, whereas S10 (MARY TTS) which is also SS-based spoofing attack present in the unknown attack was the most challenging task to detect.

These result of S10 attack degrades the performance significantly for unknown attacks. The dominance in EER for the unknown attacks was because of the S10 attack (i.e., USS based spoofing attack) than the performance of known attacks. With ESA-IFCC feature set 6.79 % was the overall average error rate and it relatively performed better than the MFCC (9.15 %) feature set. To explore the complementary information of two features. i.e., ESA-IFCC and MFCC we used score-level fusion of these two feature set. With fusion factor of $\alpha = 0.8$, and when performed score-level fusion it gave the overall average EER of 7.16 % due to dominance of S10 in unknown attack. However, the spoofing attacks from S1 to S9 present in both known and unknown attacks were detected reasonably well with the scores obtained after score-level fusion of MFCC and ESA-IFCC feature sets. The results with proposed ESA-IFCC feature set were not as good as MFCC to detect the spoofing algorithm from S1-S9. However, the proposed feature set changed the EER with only focused on S10 attack. The EER with MFCC on S10 alone was 52.99 % and for ESA-IFCC it was 37.37 %, that changes the overall average of the evaluation set and indeed obtain less EER than the MFCC even when fused at score-level.

V. SUMMARY AND CONCLUSIONS

This study presents the classification of natural *vs.* spoofed speech signal, with the components of amplitude envelope and instantaneous frequency. The modulation of AM-FM obtained with TEO-based ESA was computed with mean and standard deviation, respectively. Standard deviation of IF indeed shows the difference for classification than that of AE. Furthermore, the use of IF was made to extract the feature set. The parameters of the filterbank, namely, the number of channels, the shape of subband filters, the bandwidth, etc affects the extraction of IF components. In the future, we plan to investigate the effect of filter bandwidths. The wrong choice of filter bandwidth results in the exclusion or inclusion of modulations in the neighboring formants. The results indicates that there should be the optimum choice of bandwidth and it should be linearly-scaled.

VI. Acknowledgments

The authors would like to thank University Grants Commission (UGC) for providing Rajiv Gandhi National Fellowship (RGNF) and authorities of DA-IICT Gandhinagar.

References

[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

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

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

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

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

[6] J. Lindberg, M. Blomberg *et al.*, "Vulnerability in speaker verification-a study of technical impostor techniques." in *EUROSPEECH*, vol. 99, Budapest, Hungary, 1999, pp. 1211–1214.

[7] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA workshop*, Vigo, Spain, 2010, pp. 131–134.

[8] Villalba, Jesús and Lleida, Eduardo, "Detecting replay attacks from far-field recordings on speaker verification systems," in *European Workshop on Biometrics and Identity Management*. Springer, 2011, pp. 274–285.

[9] Y. W. Lau, D. Tran, and M. Wagner, "Testing voice mimicry with the yoho speaker verification corpus," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2005, pp. 15–21.

[10] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system." in *INTERSPEECH*, Aalborg, Denmark, 2001, pp. 759–762.

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

[12] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Kyoto, Japan: IEEE, 2012, pp. 4401–4404.

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

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

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

[16] Todisco, Massimiliano and Delgado, Héctor and Evans, Nicholas, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, 2017.

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

[18] M. R. Kamble and H. A. Patil, "Novel energy separation based instantaneous frequency features for spoof speech detection," in *European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, 2017, pp. 116–120.

[19] Madhu R. Kamble and Patil, Hemant A, "Effectiveness of mel scale-based ESA-IFCC features for classification of natural vs. spoofed speech," in *B.U. Shankar et. al. (Eds.) PReMI, Lecture Notes in Computer Sciance (LNCS)*. Springer, 2017, pp. 308–316.

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

[21] L. D. Alsteris and K. K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication*, vol. 48, no. 6, pp. 727–736, 2006.

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

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

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

[25] Maragos, Petros and Kaiser, James F and Quatieri, Thomas F, "On separating amplitude from frequency modulations using energy operators," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, San Francisco, California, USA, 1992, pp. 1–4.

[26] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *The Journal of the Acoustical Society of America (JASA)*, vol. 99, no. 6, pp. 3795–3806, 1996.

[27] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Pearson Education India, 2006.

[28] A. Potamianos and P. Maragos, "Speech analysis and synthesis using an AM–FM modulation model," *Speech communication*, vol. 28, no. 3, pp. 195–209, 1999.