Energy Separation Based Instantaneous Frequency Estimation from Quadrature and In-Phase Components for Replay Spoof Detection

Priyanka Gupta, Piyushkumar K. Chodingala, and Hemant A. Patil
Speech Research Lab, DA-IICT Gandhinagar, Gujarat, India
Email: {priyanka_gupta, piyush_chodingala, hemant_patil}@daiict.ac.in

***Abstract*—For replay Spoof Speech Detection (SSD), features that incorporate auditory transform-based information as well as Instantaneous Frequency (IF) information have been proposed in the past. IF is estimated either by derivative of analytic phase via Hilbert transform, or by using high temporal resolution Teager Energy Operator (TEO)-based Energy Separation Algorithm (ESA). However, the excellent temporal resolution of ESA comes with lacking in using relative phase information, and vice-versa. Hence, we propose novel CFCCIF-QESA features, with excellent temporal resolution as well as relative phase information. CFCCIF-QESA is designed by exploiting relative phase shift, without estimating phase explicitly. Effectiveness of proposed approach is validated by mutual information and Kullback-Leibler (KL) divergence-based analysis. Furthermore, TEO is used for complex signals for SSD. Consequently, the novel ideas of quadrature relative phase and TEO for complex signals are exploited for improving the performance of CFCCIF-ESA on ASVspoof 2017 version2.0 and ASVspoof 2019 databases. On ASVspoof 2017 evaluation set, when compared with CQCC features, CFCCIF-QESA features yield percentage improvement of 35.51% and 30.19%, with GMM and CNN classifiers, respectively. As compared to CFCCIF-ESA, a percentage improvement of 30.40% is achieved on ASVspoof 2019 evaluation dataset with GMM. Finally, the analysis of latency period indicates relatively better performance of CFCCIF-QESA and thus, its potential for practical SSD system deployment.**

***Index Terms*—Quadrature phase, Mutual Information, KL divergence, Energy Separation Algorithm, Instantaneous Frequency.**

I. Introduction

An Automatic Speaker Verification (ASV) systems is used to accept or reject the claimed speaker's identity. Recent advancements in Artificial Intelligence (AI) have led to robust and high performing ASV systems. However, ASV systems are also vulnerable to various spoofing attacks, such as impersonation by twins [1], Speech Synthesis (SS) [2], Voice Conversion (VC) [3], and replay [4]. To that effect, various challenges for Spoofed Speech Detection (SSD) have been organized in the past during INTERSPEECH conferences, such as ASVspoof 2015 [5], ASVspoof 2017 [6], ASVspoof 2019 [7], and ASVspoof 2021 [8] to develop countermeasures against spoofing attacks on ASV systems. However, out of all the known spoofing attacks, replay attacks are the easiest to mount but difficult to detect due to the availability of high-quality recording and playback devices [9].

For SSD task, in [10], an Auditory Transform (AT)-based Cochlear Filter Cepstral Coefficients-based Instantaneous Frequency (CFCCIF) feature set was proposed. It was based on cochlear filter and IF-based information. To that effect, IF is estimated conventionally from the analytic phase denoted via the Hilbert transform (HT) of the underlying real signal [11]. However, estimating IF from this approach is computationally expensive. Moreover, the resolution of HT in time-domain is poor, as it requires a block (frame) of speech data [12]. To address this issue, in [13] the authors proposed CFCCIF-ESA feature set which uses Teager Energy Operator (TEO)-based Energy Separation Algorithm (ESA) [14] to estimate IF with high time resolution for replay SSD task [15]. Due to the use of TEO in estimation of IF, CFCCIF-ESA utilizes only the amplitude information of the signal for replay SSD. Moreover, due to absence of HT, it does not contain the quadrature-phase component of the signal. Therefore, in order to incorporate both the advantages, i.e., excellent time resolution of TEO and having quadrature-phase component via HT, we propose CFCCIF-QESA feature set. Here, the term QESA represents Quadrature-based ESA. Furthermore, QESA is based on the extended definition of TEO for complex signals. To our knowledge, this extended definition of TEO is exploited for the first time for SSD task.

Additionally, the choice of quadrature-phase (90°) component along with in-phase component is justified by Mutual Information (MI)-based analysis, described in further detail in Section 2. As a result, we have developed CFCCIF-QESA feature set. To further analyze the effectiveness of considering quadrature-phase component, model-level analysis is done on Gaussian mixtures of the genuine *vs.* spoof class for both CFCCIF-ESA and CFCCIF-QESA. To that effect, Kullback–Leibler (KL) divergence between genuine and spoof Gaussian mixtures is estimated. Lastly, we also analyze the latency period of the CFCCIF-QESA in order to investigate its potential for practical SSD system deployment ability.

II. Proposed Approach

A. Exploiting Relative Phase-Based Information

So far, most of the features have been derived from the magnitude spectrum of the speech signal [16]. However, the phase characteristics can also be useful for many applications [17]–[20]. In this work, we employ an information-theoretic, approach to measure *relative* phase-based information, without estimating phase explicitly. In particular, we use Mutual Information (MI) to analyze the amount of information between

the signal and its corresponding phase-shifted signal. MI of two signals is a measure of dependence of the signals on each other, i.e., a measure of how much information the two signals share. For example, if two signals X and Y are independent, then knowing X does not yield any information about Y and vice-versa, so their MI is zero. MI is estimated as [21]:

$$I(X;Y) = h(X) - h(X|Y), \tag{1}$$

where h denotes the entropy (i.e., measure of randomness). Using the joint and marginal *pdfs* of X and Y, the MI is [21]:

$$I(X;Y) = \int_x \int_y f_{XY}(x,y) \log_2 \left(\frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} \right) dydx. \tag{2}$$

Given that speech signal can be modelled as an AM-FM signal, we consider an AM-FM signal as:

$$\begin{aligned} a(t) &= (1 + 0.5cos\,(60\pi t)), \\ x(t) &= a(t) \cos\left(2\pi f_c t + 4sin\left(2\pi f_c t + \left(\frac{\pi}{4}\right)\right)\right). \end{aligned} \tag{3}$$

For this AM-FM signal (expressed via eq. (3)) and its phase-shifted version, we have estimated the MI. The angle at which MI is minimum is the *optimum* phase value. From the values of MI obtained (as shown in Fig. 1), it can be observed that the optimum phase difference is 90° with MI=1.4349 bits. In addition, for a signal $x(t)$, the Fourier transform is denoted as $X(\omega) = X_R(\omega) + jX_I(\omega)$. Therefore, $\measuredangle X(\omega)$ is given as

$$\measuredangle X(\omega) = tan^{-1}\left(\frac{X_I(\omega)}{X_R(\omega)}\right). \tag{4}$$

From eq. (4), it can be observed that the Fourier transform phase $\measuredangle X(\omega)$ is always zero for $X_I(\omega) = 0$ which means if we do not use $\pi/2$-shifted version of $\cos(\omega t)$ (i.e., $\sin(\omega t)$) as an additional basis function in the definition of Fourier transform, it is not possible to compute $\measuredangle X(\omega)$. In this regard, Fig. 1 (b) shows the MI obtained between a cosine and its phase-shifted versions. Notably, for the cosine signal as well, MI is observed to be minimum at $\pi/2$ phase shift in $cos(\omega t)$ (i.e., $\sin(\omega t)$) indicating significance of $\cos(\omega t)$ (i.e., in phase) and its quadrature component (i.e., $\sin(\omega t)$) in the original definition of the Fourier transform. To that effect,

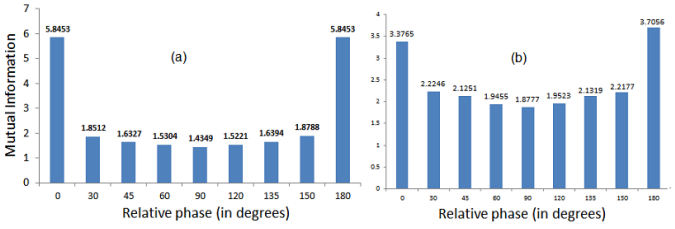

Fig. 1: MI variation w.r.t. relative phase of (a) an AM-FM signal, (b) a cosine signal and its phase-shifted version.

taking phase-shift as 90° (i.e., a quadrature) we propose an improved relative phase-based CFCCIF-QESA feature set. The feature extraction procedure of CFCCIF-QESA is shown in Algorithm 1. The quadrature component of the real-valued speech signal is achieved using Hilbert transform, which results in a complex-valued analytic signal, having a *causal* spectrum. Subsequently, TEO for complex signals is used for estimating IF using ESA. In the next sub-Section, we present the extended definition of TEO for complex signals, which is further used in the CFCCIF-QESA feature extraction procedure.

B. Extracting TEO-Based Energy for Complex Signals

As discussed above, we exploit quadrature phase-shift by estimating analytic signal. Here, we discuss the extended definition of the TEO on a complex-valued signal $z(t)$, i.e., $\psi_c[z(t)]$ which is given by [22]:

$$\psi_c[z(t)] = z(t)\dot{z}^*(t) - \frac{1}{2}[\ddot{z}(t)z^*(t) + z(t)\ddot{z}^*(t)]. \tag{5}$$

Given that $z(t)$ is complex, the TEO defined in eq. (5) is applied on real and imaginary parts of $z(t)$ separately as [23]:

$$\psi_c[z(t)] = \psi[z_r(t)] + \psi[z_i(t)]. \tag{6}$$

In this work, we extract TEO-based energy using eq. (6) on complex-valued analytic signal for improved estimation of energy as a part of ESA, discussed in the next sub-Section.

C. CFCCIF-QESA Feature Extraction

The proposed CFCCIF-QESA feature set consists of various sub-systems, as shown in Fig. 2. The filterbank of the CFCCIF-QESA consists of AT-based cochlear filters, which represent the human auditory system consisting of Basilar Membrane (BM). As per place theory of hearing [24], only a particular region of the BM vibrate in response to a particular frequency region in the speech signal. The inner hair cells act as transducers, converting the vibrations of the BM to energy. Given that the motion of the hair cell is only in the *positive* direction, it is expressed mathematically as:

$$H(a,b) = (F(a,b))^2, \tag{7}$$

where $F(a,b)$ is the output of the filterbank, and a and b govern the *size* and *shape* of each cochlear filter. The hair cell output of each filterbank is converted into a representation of the nerve spike density, which is computed as an average of $H(a,b)$ [10]. Furthermore, the quadrature-phase component

Algorithm 1: IF estimation using Quadrature-based Energy Separation Algorithm (QESA)

Input: Input: Subband filter output $f[n]$
Output: Output: IF
$f_z[n] = f[n] + j.HT\{f[n]\}$
/* Using Equation(6) */
$E_r[n] \leftarrow$ TEO{real($f_z[n]$)}
$E_i[n] \leftarrow$ TEO{imag($f_z[n]$)}
$\psi\{f_z[n]\} = E_r[n] + E_i[n]$
$IF \leftarrow Cos^{-1}\left[\frac{1-\psi\{f_z[n]-f_z[n-1]\}}{2\psi\{f_z[n]\}}\right]$

in the output $f[n]$ of the filterbank is introduced by taking its analytic signal, $f_z[n]$. This is because the analytic signal is

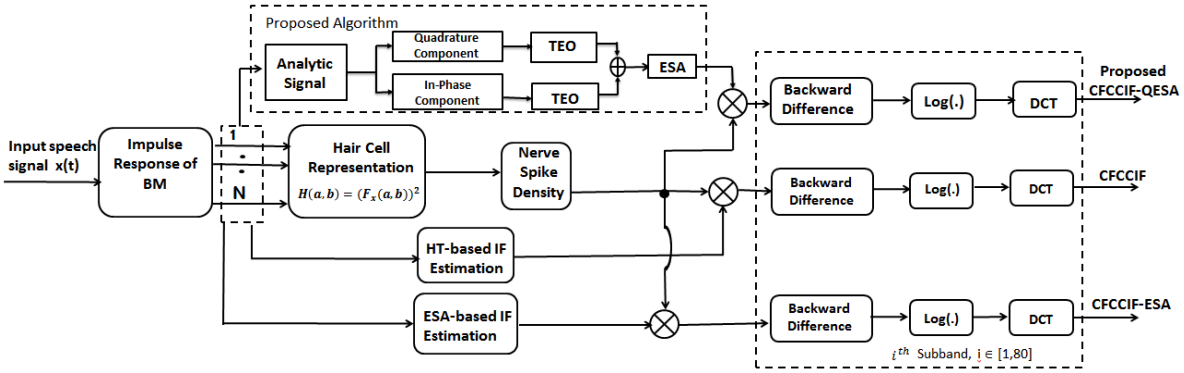

Fig. 2: Functional block diagram of proposed CFCCIF-QESA feature set (proposed algorithm denoted via dotted box), along with previous IF estimation methods using Hilbert transform and ESA to derive CFCCIF and CFCCIF-ESA, respectively.

generated by taking the Hilbert transform, which is nothing but the quadrature-shifted version of $f[n]$. Now, in order to estimate the energy of the complex-valued analytic signal, we use the extended definition of TEO as described in Section II-B. Furthermore, the energy profile obtained from the extended TEO is used to estimate IF using ESA. The ESA for IF estimation of a signal $f(n)$ is given by [24]:

$$\omega_i[n] \approx Cos^{-1}\Big[\frac{1-\psi\{f[n]-f[n-1]\}}{2\psi\{f[n]\}}\Big]. \qquad (8)$$

Here, the $\psi\{f[n]\}$ represents the Teager energy of $f[n]$, and the $\omega_i[n]$ represents the estimated IF.

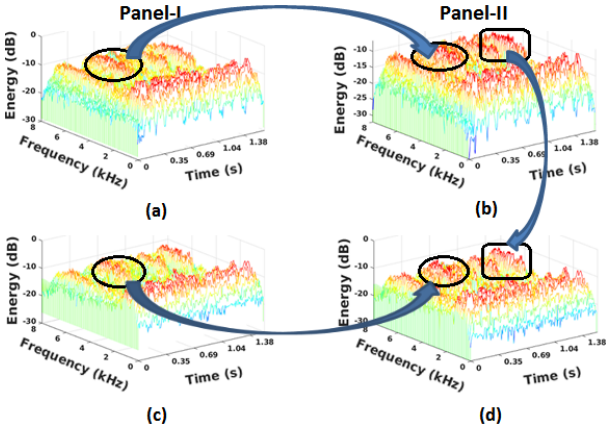

Fig. 3: Panel I and Panel II show waterfall plots for CFCCIF-ESA and CFCCIF-QESA, respectively. Here, (a) and (b) are corresponding to genuine speech signal, and (c) and (d) are corresponding to spoofed speech signal.

D. Model-level Measure of *CFCCIF-ESA* vs. *CFCCIF-QESA*

In order to show the efficiency of the proposed CFCCIF-QESA, we estimate the model-level measure of CFCCIF-ESA (i.e., without quadrature-phase component) and CFCCIF-QESA (i.e., with quadrature-phase component) on ASVspoof 2017 and ASVspoof 2019 datasets. KL divergence between statistical Gaussian Mixture Model (GMM) of genuine and spoofed speech, is used as a model-level measure of discriminative ability [25]. To that effect, we have estimated KL divergence between genuine and spoof GMMs of the two

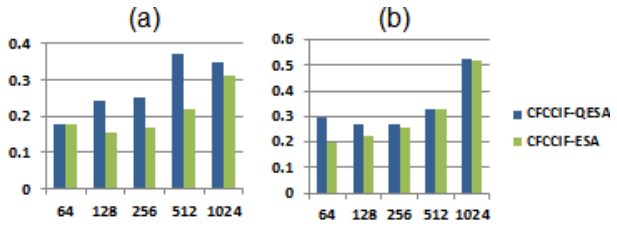

Fig. 4: (a) KL divergence between genuine and spoof Gaussian mixtures on (a) ASVspoof 2017, and (b) ASVspoof 2019

feature sets CFCCIF-ESA and CFCCIF-QESA. A higher value of KL divergence can indicate better discriminating ability of the GMM.

KL divergence tells us about how much one discrete probability distribution function (PDF) differs from a second PDF. For SSD task, it has been used as a model-level measure for distinguishing genuine *vs.* spoof class [26]. If p and q are two discrete PDFs, then it is a measure of the information lost, when $q(x)$ is used to approximate $p(x)$. Mathematically, it is expressed as [25]:

$$KL(p||q) = -\int_x p(x) \ln\Big\{\frac{q(x)}{p(x)}\Big\}\, dx. \qquad (9)$$

From Fig. 4 it is worth noting that the KL divergence between genuine and spoof Gaussian mixtures is higher for CFCCIF-QESA than that for CFCCIF-ESA features, indicating better discriminative ability of proposed feature set which is shown by the waterfall plots in Figure 3 and is also reflected in results, discussed in the next Section.

III. Experimental Results

A. Datasets Used

In this study, we have used ASVSpoof 2017 V2.0 (real replay) and ASVSpoof 2019 (Physical Access via simulated replay) corpora, having 16 kHz sampling frequency. The statistics of both the datasets are given in [6], [7].

B. Features and Classifier Used

The experiments are performed using auditory transform-based features, namely, CFCCIF, CFCCIF-ESA, and CFCCIF-QESA, with the baseline CQCC features. CFCCIF-QESA is implemented using 80 linearly-spaced filterbanks, with the filter parameters as $a = 3$ and $b = 0.016$, which have been

TABLE I: Results (in % EER) using GMM and CNN on ASVspoof 2017 v2.0 and GMM on ASVspoof 2019 PA datasets

Feature Set ↓	**Dev. (GMM)**	**Eval. (GMM)**	**Dev. (CNN)**	**Eval. (CNN)**	**Dev. (GMM)**	**Eval. (GMM)**
Dataset →	**ASVspoof 2017 V2.0**				**ASVspoof 2019 PA**	
CQCC(baseline B1)	12.87	18.81	10.00	28.42	9.87	11.04
CFCCIF (S1)	16.61	17.38	10.92	16.40	36.93	37.61
CFCCIF-ESA (S2)	11.54	14.77	11.31	19.02	36.29	36.94
Quadrature based CFCCIF-ESA (S3)	**9.71**	**12.13**	7.67	14.00	22.39	25.71
B1⊕S3	9.54	12.09	**2.19**	**11.00**	**9.87**	**11.01**
S1⊕S3	9.6	11.98	7.13	12.80	22.20	25.60
S2⊕S3	9.52	11.94	6.55	13.33	22.18	25.58
S1⊕S2⊕S3	**9.51**	**11.93**	6.37	12.86	22.38	25.65

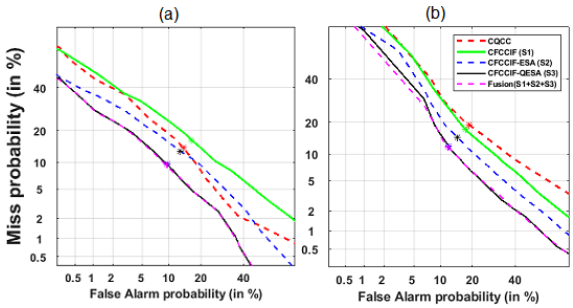

Fig. 5: DET curves for the systems on (a) development set (b) evaluation set of ASVspoof 2017 dataset using GMM.

found empirically for optimal performance. All the feature sets are *36*-dimensional with static, Δ, and $\Delta\Delta$ coefficients. Furthermore, to enhance the performance, Cepstral Mean and Variance Normalization (CMVN) is applied on feature sets, which eliminates the channel distortion [27].

For classification task, two classifiers, namely, GMM and Convolutional Neural Network (CNN) have been used. The GMM is used, with 512 mixture components, to train the model. The performance of the model is evaluated using %EER. The CNN consists of 5 convolution blocks and 4 Fully-Connected (FC) layers. Each convolution block includes a *2*-D convolution layer followed by batch normalization and max-pooling layer. The operations in both convolution and max-pooling layers are done using kernels of sizes 3×3 and 2×2, respectively. In the convolution layer, the padding and stride are kept at 1. The size of input to the first convolution layer is 36×400. The batch size and learning rate are set as 32 and 0.0001, respectively. The activation function used is the Rectified Linear Unit (ReLU). The cross-entropy loss is used as the loss function.

C. Results

The obtained experimental results (shown in Table I) on the proposed CFCCIF-QESA features are compared with the existing auditory transform-based features, such as CFCCIF and CFCCIF-ESA. The results are also compared with the ASVspoof 2017 challenge baseline CQCC features. It can be observed that the CFCCIF-QESA feature set outperform all the other features on ASVspoof 2017 v2.0 dataset, for both GMM and CNN classifiers. Further on ASVspoof 2017 dataset, Fig. 5 shows the Detection Error Trade-off (DET) curves for all the systems, including their fusion systems. On the evaluation set of ASVspoof 2017 dataset, as compared to CFCCIF-ESA (non-quadrature based), CFCCIF-QESA gave a percentage improvement of 17.87% and 22.59% with GMM and CNN, respectively. Furthermore, the score-level fusion of the proposed CFCCIF-QESA feature set with CFCCIF and CFCCIF-ESA yields further reduced $\%$ EER in all the cases. In order to capture the complementary information of each system, score-level fusion is done on Log-Likelihood Ratio (LLR) scores, by computing the *linear weighted sum* as

$$LLR_{fused} = \alpha \cdot LLR_{feat1} + (1 - \alpha) \cdot LLR_{feat2}, \quad (10)$$

where LLR_{feat1} and LLR_{feat2} are LLR scores derived from feature set-1 and feature set-2, respectively. The Table I shows the score-level fusion of the auditory transform-based systems $S1, S2, S3$, where $\oplus$ denotes fusion. The contribution of each feature set in the scores obtained from fusion is determined by $\alpha \in [0, 1]$. On ASVspoof 2017 dataset, values of α for systems $S1 \oplus S2$, $S2 \oplus S3$, and $S1 \oplus S2 \oplus S3$ are $0.95, 0.093$, and 0, respectively, for development, and $0.67, 0.75$, and 0.08, respectively, for evaluation set.

D. Analysis of Latency Period

Latency period represents the performance evaluation in terms of %EER *w.r.t* different durations of speech segment in an utterance. The utterance duration ranges from 20 ms to 2 seconds, with an interval of 200 ms. Further, the utterance duration is selected by considering the number of frames. Figure 6 shows comparison between the CQCC baseline, CFCCIF, CFCCIF-ESA, and CFCCIF-QESA. It can be observed that all the feature sets show comparable latency with each other for the development set of ASVSpoof 2017 as shown in Figure 6. However, for the evaluation set of ASVSpoof 2017 as shown in Figure 6 (b), we observe a considerable improvement of CFCCIF, CFCCIF-ESA, and CFCCIF-QESA in latency performance w.r.t. CQCC baseline. Furthermore, the %EER converges to the minimum value as the speech duration provided to the model of SSD system increases. Additionally, the feature performance is better if for a low latency period the $\%EER$ is also low, indicating faster classification by the model and thus, indicating suitability of the system for practical deployment.

IV. Summary And Conclusions

In this study, QESA is proposed for the first time for utilizing information captured by relative phase-shift between signals as well as exploiting excellent time resolution of ESA. To that effect, auditory transform-based CFCCIF-QESA

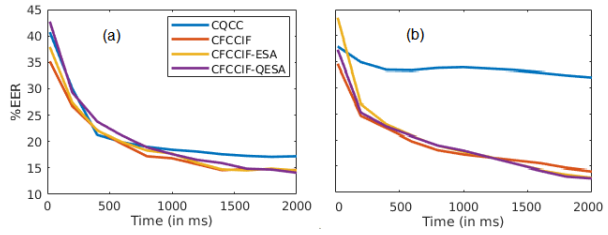

Fig. 6: Analysis of latency period for the SSD system (a) dev. set (b) eval. set of ASVspoof 2017 dataset using various feature sets.

feature set is proposed. MI-based analysis is done to determine the optimum relative phase shift. It is found that a quadrature phase shift is the best suited. Further, MI is used to justify the basis functions used in the original definition of Fourier transform. To that effect, the signal is converted to its analytic signal (which has its real and imaginary parts separated by a quadrature phase). The analytic signal is complex-valued and hence, for the first time, the extended definition of TEO for complex signals is used for the SSD task. Experiments are performed on ASVspoof 2017 version 2.0 and ASVspoof 2019 datasets and CFCCIF-QESA features are shown to perform better than features without quadrature-phase on ASVspoof 2017 dataset using GMM. However, the limitation of this work is that the proposed features does not yield improved performance than CQCC baseline for ASVspoof 2019 PA dataset (which is in agreement with recent findings in [28]), because it contains simulated replay utterances, unlike ASVspoof 2017 dataset which contains replay utterances under realistic scenarios. Our future efforts will be directed towards investigate performance of CFCCIF-QESA for SS and VC-based attacks on ASV.

V. Acknowledgements

The authors would like to thank the Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, for sponsoring consortium project titled 'Speech Technologies in Indian Languages' under 'National Language Translation Mission (NLTM): BHASHINI', subtitled 'Building Assistive Speech Technologies for the Challenged' (Grant ID: 11(1)2022-HCC (TDIL)). We also thank the consortium leaders Prof. Hema A. Murthy, Prof. S. Umesh, and the authorities of DA-IICT Gandhinagar, India for their support and cooperation to carry out this research work.

References

[1] "HSBC reports high trust levels in biometric tech as twins spoof its voice ID system," *Biometric Technology Today*, vol. 2017, no. 6, 2017.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, 2009.

[3] Y. Stylianou, "Voice transformation: A survey," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 19-24 April, 2009, pp. 3585–3588.

[4] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection," *INTERSPEECH 2017*, Stockholm, Sweden, 20-24 August 2017.

[5] W. Zhizheng, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Sixteenth Annual Conference of the International Speech Communication Association*, Dresden, Germany, 6-10 September, 2015, pp. 2037–2041.

[6] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVSpoof 2017 Version 2.0: Meta-Data Analysis and Baseline Enhancements," in *Odyssey- The Speaker and Language Recognition Workshop, France*, 26-29 June 2018, p. 296–303.

[7] A. et al., "ASVSpoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[8] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Trans. on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[9] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. INTERSPEECH, Lyon, France*, 25-28 August 2013, pp. 925–929.

[10] T. B. Patel and H. A. Patil, "Cochlear filter and instantaneous frequency based features for spoofed speech detection," *IEEE Journal of Selected Topics in Sig. Process.*, vol. 11, no. 4, pp. 618–631, 2016.

[11] Tanvina B. Patel and Hemant 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*, 6-10 September 2015, pp. 2062–2066.

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

[13] A. T. Patil, R. Acharya, P. K. A. Sai, and H. A. Patil, "Energy separation-based instantaneous frequency estimation for cochlear cepstral feature for replay spoof detection," in *INTERSPEECH, Graz, Austria*, 15-19 September 2019, pp. 2898–2902.

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

[15] J. F. Kaiser, "On a simple algorithm to calculate the'energy'of a signal," in *ICASSP, Albuquerque, NM, USA*, 3-6 April 1990, pp. 381–384.

[16] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.

[17] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, 1981.

[18] I. Saratxaga, I. Hernaez, M. Pucher, E. Navas, and I. Sainz, "Perceptual importance of the phase related information in speech," in *INTERSPEECH, Portland, USA*, 09 September 2012.

[19] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining mfcc and phase information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1085–1095, 2011.

[20] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Trans. on Info. Forensics and Security*, vol. 10, no. 4, pp. 810–820, 2015.

[21] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[22] R. Hamila, J. Astola, F. A. Cheikh, M. Gabbouj, and M. Renfors, "Teager Energy and the Ambiguity Function," *IEEE Transactions on Signal Processing*, vol. 47, no. 1, pp. 260–262, 1999.

[23] P. Maragos and A. C. Bovik, "Image demodulation using multidimensional energy separation," *Journal of the Optical Society of America A (JOSA)*, vol. 12, no. 9, pp. 1867–1876, 1995.

[24] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. 2^{nd} Edition, Pearson Education India, 2004.

[25] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, 2017.

[26] K. Sriskandaraja, V. Sethu, P. N. Le, and E. Ambikairajah, "Investigation of sub-band discriminative information between spoofed and genuine speech." in *INTERSPEECH, San Francisco, CA, USA*, 8-12,September2016, pp. 1710–1714.

[27] A. T. Patil and H. A. Patil, "Significance of CMVN for replay spoof detection," in *APSIPA-ASC*, 7-10 December 2020, pp. 532–537.

[28] A. T. Patil, H. A. Patil, and K. Khoria, "Effectiveness of energy separation-based instantaneous frequency estimation for cochlear cepstral features for synthetic and voice-converted spoofed speech detection," *Computer Speech & Language*, p. 101301, 2021.