Teager Energy Subband Filtered Features for Near and Far-Field Automatic Speech Recognition

Madhu R. Kamble*, Shekhar Nayak†, M. Ali Basha Shaik†, Shakti P. Rath†, Vikram Vij† and Hemant A. Patil‡
* EURECOM, Sophia Antipolis, France.
† Samsung Research and Development Institute, Bangalore, India.
‡ Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India.
madhu.kamble@eurecom.fr, {s1.nayak, m.shaik, shakti.rath, vikram.v}@samsung.com, hemant_patil@daiict.ac.in

***Abstract*—Automatic Speech Recognition (ASR) usually works well with close-talking microphone environment rather than in far-field conditions. A major challenge in the far-field ASR systems is to handle the background noise, multipath reflections, and reverberation, that leads to decrease in the quality of the speech signal. To that effect, we propose Teager energy-based Gabor filterbank (TGFB) features that preserve the amplitude and frequency modulation of a resonant signal, and improve the time-frequency resolution. In addition, via TGFB features, we exploit noise suppression capability of Teager Energy Operator (TEO) for improving ASR performance under signal degradation conditions due to far-field speech. The ASR experiments are performed on LibriSpeech (near-field) and CHiME-3 (far-field) corpora. Marginal improvements were observed for TGFB features over MFCC features in our experiments. We observed that the system combination of TGFB and MFCC features could provide significant improvements over the standalone MFCC features. For LibriSpeech corpus, a relative improvement for Word Error Rate (WER) of close to 5% was observed. On the other hand, for CHiME-3 corpus, the average relative improvement of 7.20% was obtained over the baseline features using system level combination.**

***Index Terms*—Automatic Speech Recognition, Teager Energy Operator, Near and Far-Field.**

I. Introduction

Automatic Speech Recognition (ASR) is a task that converts a speech signal into a continuous sequence of words along with real interaction between humans and the machines [1]. Far-field speech recognition is an essential technology for interactions that aims to provide access of the smart devices through the recognition of far-field speech [2]. This technology is applied to smart home appliances (smart loudspeaker, and TV), meeting transcription, and onboard navigation, etc. However, in a real environment, there is a lot of background noise, multipath reflections, reverberation, and even human voice interference, leading to decrease in ASR accuracy [3].

Recent developments in acoustic modeling employs techniques, such as deep learning, sequence modeling, etc. However, their performance degrades in the case of far-field recording conditions. The reverberant artifacts distort the speech signal by smearing the amplitude envelopes of the speech signal [4]. The development of a real-world applications faces a notable challenge because of reverberation. The ASR system degrades the performance when the far-field microphone array signals are used instead of close-talking microphone.

The aim of the 3^{rd} CHiME (Computational Hearing in Multisource Environments) challenge was to develop a multi-channel ASR system [5]. The CHiME-3 dataset upgrades the difficulty by providing not only artificially noisy speech (i.e., obtained by combining clean speech with recorded background noise) but also consists of the noisy speech recorded in public environments, such as cafe, bus, street junction, and pedestrian areas. The CHiME-3 challenge covers different speakers, noise environments, and real-world problems, such as clipping, microphone failure, recording glitches, etc.

Our goal in this work is to increase the robustness of ASR using Teager energy-based features in noise and reverberation in order to combine them efficiently with standard Mel Frequency Cepstral Coefficients (MFCC)-based front-ends with GMM (Gaussian Mixture Model), and DNN (Deep Neural Network) acoustic models in addition to use of RNN (Recurrent Neural Network) as language models. The motivation behind using Teager Energy Operator (TEO) is its attributes to capture nonlinear aspects of speech production [6]. The true total energy of source is estimated using TEO, and it also preserves the amplitude and frequency modulation of a resonant signal. Hence, it improves the time-frequency resolution along with improving the formant information representation [7]. In addition, the TEO has the noise suppression property, and it attempts to suppress the distortion caused by noise signal. While there are studies in ASR literature that exploit noise suppression capability of TEO for ASR task, however, they are either for close-talking speech [8] or artificially added noise [9]. The present study extends this work for a typical acoustic noise characteristics of real far-field scenarios.

The rest of the paper is organized as follows: Section II presents the basic mathematical details of the Teager Energy Operator (TEO). In addition, Section II also presents the block diagram of the proposed feature set, the spectral energy differences along with noise suppression capability of TEO. The experimental setup is explained in Section III along with some basic information of the corpus used. Section IV presents the experimental results on both the databases, i.e., LibriSpeech and CHiME3 corpus. Finally, Section V concludes the paper

along with future research directions.

II. Teager Energy Operator (TEO)

The TEO tracks running estimate of instantaneous energy fluctuations of a narrowband speech signal [7], [6], [10]:

$$\Psi_d\{x_i[n]\} = x_i^2[n] - x_i[n-1]x_i[n+1] \approx a_i[n]^2\Omega_i[n]^2, \quad (1)$$

where $x_i[n]$ is discrete-time bandpass filtered signal for i^{th}

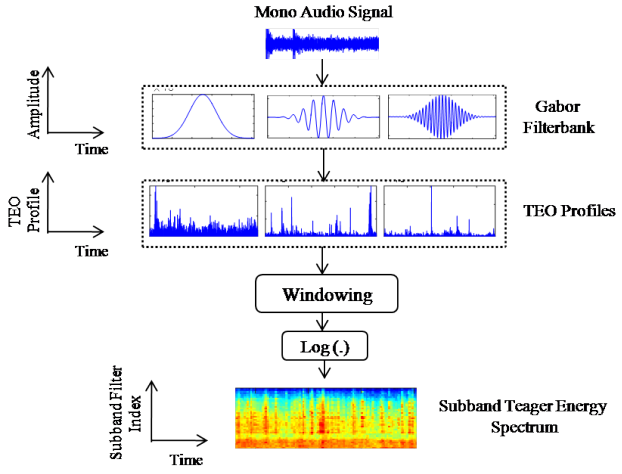

Fig. 1. Block diagram of Teager energy spectral features. After [11], [12].

subband filter, and $\Psi_d\{\cdot\}$ represents TEO. The TEO works on narrowband signal and hence, bandpass filtering is necessary to apply on the input speech signal to compute *'N'* number of subband filtered signals. The block diagram of Gabor filterbank energy coefficients-based on TEO is shown in Fig. 1. Here, the input speech signal is first passed through the Gabor filterbank to obtain *'N'* subband filtered signals [4], [13], [14]. We used Mel-spaced Gabor filterbank to have compressed bandwidth in the lower frequency region and wide bandwidth in the higher frequency regions. The optimal criteria here is to be able to achieve minimum time-bandwidth product that is dictated by Heisenberg's uncertainty principle in signal processing framework [15]. The temporal variance (σ_t^2) and the frequency variance (σ_ω^2) of a signal, $f(t) \in L^2(R)$ (i.e., Hilbert space of finite energy signals) satisfy,

$$\sigma_t^2.\sigma_\omega^2 \geq 1/4. \quad (2)$$

This inequality becomes equality if and only if $f(t)$ is Gaussian, where $\sigma_t^2.\sigma_\omega^2$ is called as time-bandwidth product (which is also area of Heisenberg box). The earlier studies found that the linearly-spaced center frequencies have good resolution in both the lower and higher frequency regions which makes the estimation of the spectral information more reliable [14]. Hence, the narrowband filtered signals are obtained at center frequency, which are Mel-spaced between f_{min}=10 Hz, and f_{max}=8000 Hz. The impulse response, $h(t)$, of Gabor filter is given by [7] :

$$h(t) = exp(-b^2t^2)cos(\omega_c t), \quad (3)$$

where ω_c is the center frequency (in Hz) of the subband filter chosen as per the Mel frequency scales. The parameter b controls the bandwidth of the subband filter. The Gabor filter has the linear phase response characteristics and hence, it maintains the same pattern (shape) of the filtered speech signal (within the passband of filter) with a delay in time which is equal to group delay function (in seconds) of filter [16].

In ASR, the lower formants and harmonics are important as the linguistic information is present in lower formant frequencies and hence, these should be preserved. Furthermore, these subband filtered signals are passed through TEO block in order to estimate the Teager energy profile of each subband filtered signals. These Teager energy profiles are further passed to the frame-blocking along with averaging of the speech segment using a window length of 25 ms and shift of 10 ms followed by logarithm operation. Finally, these filterbank energy coefficients are extracted from the speech signal. Henceforth, we will denote it as TGFB (Teager energy-based Gabor filterbank) feature set for the ASR task.

The time-frequency representations of the speech signal from the CHiME-3 corpus is shown in Fig. 2. The comparison is done with time-frequency representations obtained from the Mel filterbank, and TGFB features as shown in Fig. 2(b), and Fig. 2(c), respectively, for both real (Panel I) and simulated (Panel II) speech signals. It can be observed that the energy obtained for the real speech signal from the Mel filterbank has less energy spectral density compared to the TGFB approach. For the ASR task, the lower formants, such as F_1 and F_2 are important to preserve the linguistic content. The spectral energy obtained from the TEO shows the sharp formants, and high energy compared to that of Mel filterbank features. In particular, the Mel spectral energy obtained are distorted, and have blurred characteristics at the higher frequency regions.

A. Noise Suppression Capability of TEO

The noise suppression capability of TEO was originally analyzed for near-field speech in car noise as acoustic environment in [17] followed by our recent work on Wall Street Journal (WSJ) corpus [9]. Consider a clean speech signal $s(n)$, degraded by a additive noise $\eta(n)$, and the resulting in noisy speech signal is given as $y(n)$:

$$y(n) = s(n) + \eta(n). \quad (4)$$

The TEO of the noisy speech signal is given by:

$$\psi[y(n)] = \psi[s(n)] + \psi[\eta(n)] + 2\tilde{\psi}[s(n)\eta(n)], \quad (5)$$

where $\tilde{\psi}[s(n)\eta(n)]$ is the cross-ψ energy of $s(n)$ and $\eta(n)$. As $s(n)$ and $\eta(n)$ are statistically-independent, the expected value of $\tilde{\psi}[s(n)\eta(n)]$ is zero [8], [17] and hence,

$$E\{\psi[x(n)]\} \approx E\{\psi[s(n)]\} + E\{\psi[\eta(n)]\}. \quad (6)$$

We analyzed the power spectral density (PSD) of a speech segment for far-field data (speech signals are taken from CHiME-3 corpus). The PSD plots obtained from the (a) street, (b) pedestrian, (c) bus, and (d) cafe background environment obtained from with and without TEO (applied as speech

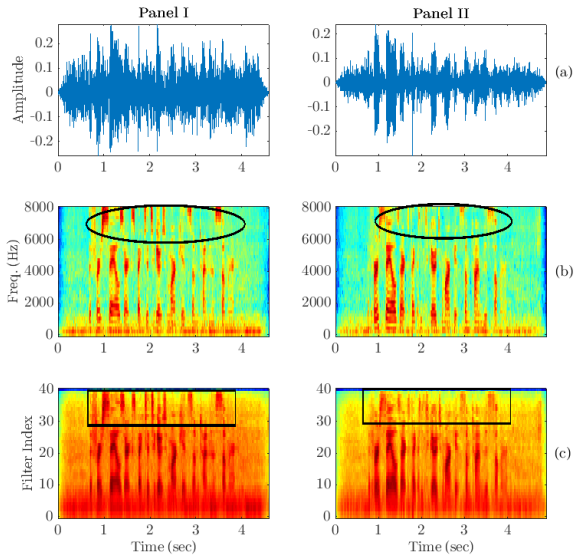

Fig. 2. (a) Time-domain speech signal for (Panel I) real, and (Panel II) simulated, spectral energy density obtained from (b) Mel filterbank, and (c) TGFB. Highlighted ovals and box shows the spectral energy differences between Fig. 2(b) and Fig. 2(c).

signals from CHiME 3 corpus) is shown in Fig. 3. Noise suppression capability of TEO can be clearly observed in Fig. 3, in particular, the PSD plot of noisy speech (shown in red color in Fig. 3) [9]. [18] shifts downward in TEO-domain indicating noise suppression that is achieved due to mathematical structure of TEO.

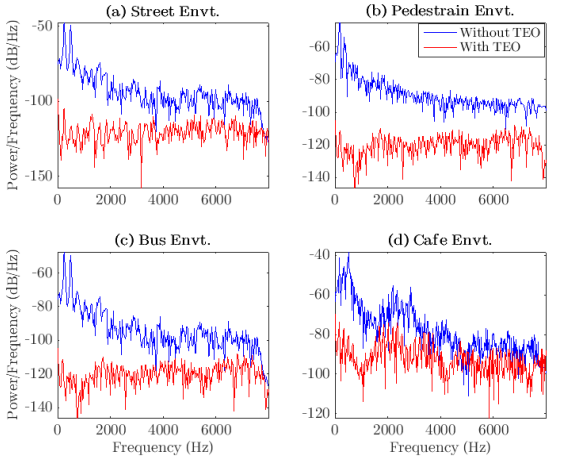

Fig. 3. Power Spectral Density (PSD) of a real speech segment with street, pedestrian, bus, and cafe background. The PSD is shown for speech segment with and without application of TEO. Downward shift in PSD plot with TEO indicates noise suppression due to TEO.

III. Experimental Setup

A. Near-Field and Far-Field ASR Corpus

In this paper, the ASR experiments were performed on LibriSpeech and CHiME3 corpora. The LibriSpeech task comprises English read speech data based on the LibriVox project [19]. The LibriSpeech database consists of two sets of clean speech data (100 hours + 360 hours), and noisy speech data (500 hours) for training. We used 100 hours of clean speech data to train the initial ASR model, and tested the trained models on test-clean and test-other subsets of Librispeech. The statistics of the database is reported in [19]. In addition, we also performed experiments on CHiME-3 corpus which uses multi-microphone tablet device in everyday environments [5]. Four varied environments have been selected: cafe (CAF), street junction (STR), public transport (BUS), and pedestrian area (PED). The real speech data is of *6*-channel recordings of the sentences from the WSJ0 (Wall Street Journal) corpus. The simulated data was developed by adding the clean speech utterances with the different environment in the background during recordings. For ASR evaluation, the corpus is divided into three subsets, namely, training, development, and test sets, respectively.

B. Feature Representation

For Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) training, MFCC features are extracted from the speech signals using a window length of 25 ms and shift of 10 ms. Delta and double-delta features are also appended resulting in *39*-dimensional (D) feature set. Human auditory system can be viewed as a dense filterbank in frequency-domain with several thousands of subband filters [18]. Hence, we performed experiments to investigate the significance of number of subband filters on the ASR performance. The TGFB features are extracted using subband filters by the process shown in Fig. 1.

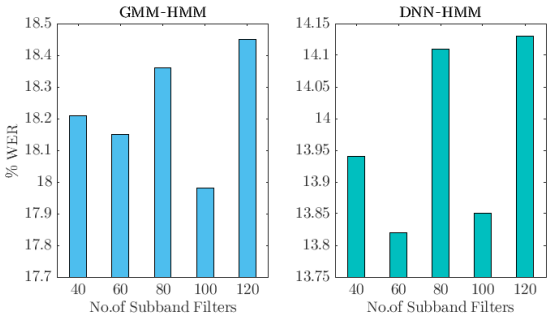

Fig. 4. Effect of subband filtered signals on TGFB feature set.

Fig. 4 shows the effect of increasing number of subband filtered signals during feature extraction. The experiments were performed with GMM-HMM and DNN-HMM systems by varying the number of subband filters from 40 to 120. It can be observed from the Fig. 4 that features extracted using 60 number of subband filters are found to be *optimal* on DNN-HMM systems compared to the other number of subband

TABLE I
WER (%) USING BEAMFORMING AND ENHANCED METHODS WITH PROPOSED FEATURE SET TRAINED ON MULTI-ENHANCED SPEECH

Method	Dev				Eval					
	Real		Sim		Real		Sim		Avg.(Real+Sim)	
	MFCC	TGFB	MFCC	TGFB	MFCC	TGFB	MFCC	TGFB	MFCC	TGFB
GMM-HMM	16.59	16.96	18.93	19.35	26.55	26.25	26.73	25.71	26.64	**25.98**
+ DNN (CE)	13.19	12.95	14.66	14.69	20.76	20.30	20.75	19.79	22.25	**20.04**
+ DNN (sMBR)	11.73	12.06	13.26	13.53	18.63	18.53	18.82	18.32	18.72	**18.42**
+ 5-gram rescoring	10.59	10.81	11.85	12.01	17.01	16.68	16.95	16.60	16.98	**16.64**
+ RNNLM	9.86	9.89	11.18	11.41	15.97	15.61	15.67	15.47	15.82	**15.54**

TABLE II
WER(%) FOR EACH NOISE WITH MFCC AND TGFB FEATURES AND WITH THE SYSTEM-LEVEL COMBINATION (SC) OF MFCC AND TGFB FEATURE SET

	MFCC				TGFB				SC			
Envt.	Dev		Eval		Dev		Eval		Dev		Eval	
	Real	Sim	Real	Sim	Real	Sim	Real	Sim	Real	Sim	Real	Sim
Avg.	13.19	14.66	20.76	20.75	9.89	11.41	15.61	15.47	**9.43**	**10.79**	**14.78**	**14.59**
BUS	15.96	13.6	27.43	15.63	11.71	10.41	21.63	12.10	**11.37**	**9.96**	**20.56**	**11.11**
CAF	12.82	16.52	19.95	21.31	9.29	13.75	13.37	16.01	**8.70**	**12.58**	**12.29**	**15.11**
PED	10.43	12.70	18.59	22.69	8.16	9.59	13.66	16.16	**7.85**	**9.14**	**12.76**	**15.28**
STR	13.54	15.77	17.07	23.35	10.41	11.89	13.77	17.63	**9.79**	**11.46**	**13.50**	**16.87**

filtered signals and hence, further set of experiments for TGFB features were performed with 60 number of subband filtered signals. In this papaer, the Kaldi toolkit is used to build the ASR systems for both the corpora [20].

IV. EXPERIMENTAL RESULTS

A. Results for Near-Field ASR

GMM-HMM system is used to generate the alignments for training the DNN-based model and also as the baseline system. MFCC features (*39*-dimensional) are spliced with 7 frames context and linear discriminant analysis is applied to project it to *40*-dimensions and further semi-tied covariance is applied. Then, speaker adaptive training using a single feature-space maximum likelihood linear regression (FMLLR) transform is done to obtain the resultant features for training the GMM-HMM systems [21]. The MFCC-based GMM-HMM baseline system achieved WERs of 12.28 %, and 34.92 % on test-clean and test-other portions of Librispeech, respectively. The results obtained with TGFB are comparable to the MFCC feature set resulting in 12.71 % and 35.58 % WER on test-clean and test-other, respectively.

The experiments performed consists of DNN with 6 hidden layers with 1024 neurons (sigmoid activation) in each hidden layer. The output units are 3480 senones or context-dependent triphones for each DNN which are obtained using forced alignment from the GMM-HMM system. The input is 11 frames (5 left context, 1 current, and 5 right context) of *60*-D features concatenated together. The performance of the ASR system is analyzed using Word Error Rate (WER). The DNN-HMM system is trained on clean speech and tested for both test-clean and test-other for MFCC and TGFB feature sets. The experimental results of test set using the DNN-HMM systems are reported in Table III. This shows that TGFB features are as effective as MFCC features for ASR in near-field conditions.

TABLE III
WER (%) ON DNN-HMM SYSTEM TRAINED ON 100 HRS OF TRAINING DATA OF LIBRISPEECH CORPUS

Subset	# Hrs	MFCC	TGFB
Test-clean	5.4	9.55	**9.40**
Test-other	5.1	27.62	**27.54**

B. Results for Far-Field ASR

The speech enhancement baseline system based on time-varying minimum variance distortionless response (MVDR) beamforming available in Kaldi is used for transforming the multi-channel noisy input signals to single channel enhanced output signals [5]. The training set used is clean speech taken from WSJ0 corpus and multi-noisy data and tested on noisy speech signals. On the other hand, the enhanced speech signal were tested on the clean speech and multi-enhanced speech signal. The DNN has 7 layers with 2048 units per hidden layer. The input layer has 5 frames of left and right context (i.e., 11×40 = 440 units). The DNN is pre-trained using restricted Boltzmann machines, cross-entropy (CE) training, and sequence discriminative training using the state-level minimum Bayes risk (sMBR) criterion. In addition, the *N*-gram rescoring, and RNN-based LM (RNNLM) is used for far-field ASR task. The experimental results with GMM, DNN, and RNN-LM-based ASR system are shown in Table I which is trained on multi-enhanced speech signal with MFCC and TGFB feature sets.

Furthermore, in order to combine the possible complementary advantages available from the amplitude and frequency modulation of the speech signal, the posterior lattices obtained

TABLE IV
SPEECH RECOGNITION PERFORMANCE (IN % WER) ON NEAR-FIELD AND FAR-FIELD CORPUS

Corpus	System	Subset	SC	RI (%)
LibriSpeech	Near-field	Test-Clean	**9.15**	**4.19**
LibriSpeech	Near-field	Test-Other	**26.24**	**4.99**
CHiME-3	Far-field	-	**14.68**	**7.20**

(SC: System Combination, RI: Relative Improvement)

from the MFCC and TGFB feature sets were combined using lattice-level system combination (as shown in Table IV) [22], [23]. The performance of the combined system is 9.15% and 26.24% for test-clean and test-other, respectively, on LibriSpeech corpus, that results in relative improvement of 4.19% and 4.99% which is better than the MFCC alone. On the other hand, for CHiME-3 corpus, a relative improvement of 7.20% is obtained resulting in 14.68% WER.

The detailed results on different noises in CHiME-3 are reported in Table II with MFCC, TGFB features and their system combination. For all the noise conditions of CHiME-3 corpus, the TGFB feature set shows improvements over the baseline system on the evaluation set. This shows the noise supression capability of TEO in varied noisy conditions. The individual noise condition performance after system combination (SC) of MFCC and TGFB using MBR decoding significantly improved over the standalone features.

TABLE V
COMPARISON WITH OTHER SYSTEMS ON CHIME-3 CORPUS

System	Dev		Eval	
	Real	Sim	Real	Sim
MFB [24]	11.6	14.3	22.6	25.5
PFB [24]	12.0	13.7	23.0	25.1
RAS [24]	11.8	14.6	21.6	23.1
MHE [24]	12.0	14.4	22.9	26.4
CVAE[24]	10.2	12.4	18.9	19.9
Ratemap+F_0 [25]	5.51	4.82	18.56	20.03
PNCC [26]	14.23	11.85	22.12	14.88
MESSL [27]	9.00	11.5	16.3	21.00
log Mel [28]	12.58	10.66	23.86	20.17
DOC [28]	12.00	10.18	20.35	18.53
NIN-CNN [29]	10.64	11.21	12.81	18.47
DS Beamforming [30]	13.92	13.62	26.30	21.14
MFCC+RNNLM	9.86	11.18	15.97	15.67
TGFB	12.06	13.53	18.53	18.32
TGFB+RNNLM	9.89	11.41	15.61	15.47
TGFB+RNNLM_SC	9.43	10.79	**14.78**	**14.59**

Finally, we compared the performance of TGFB feature set with our proposed and the other similar systems (in the literature) as reported in Table V. Currently, we have not considered end-to-end speech recognition systems for evaluation and comparison purpose as the purpose of this work is majorly to highlight the importance of noise robust Teager energy-based features for ASR.

V. SUMMARY AND CONCLUSIONS

In this study, we presented the use of Teager energy spectral features-based acoustic model for near and far-field ASR tasks, where the TGFB feature set was extracted from Mel-spaced Gabor filterbank. The TEO preserves the amplitude and frequency modulation of a resonant signal, and it improves the time-frequency resolution. The noise suppression capability of TEO indeed helps for robust ASR task. The performance of the ASR system degrades when far-field speech is considered instead of near-field speech and hence, far-field system are more challenging, in particular, to handle background noise, reverberation, etc. The experiments are performed on both LibriSpeech (near-field) and CHiME-3 (far-field) corpora. Significant reduction in % WER was achieved using the system combination using MBR decoding of MFCC and TGFB-based features. For LibriSpeech corpus, we obtained relative improvement of 4.19% and 4.99% in word error rate (WER) for test-clean and test-other, respectively. On the other hand, for CHiME-3 corpus, the average relative improvement of 7.20% is obtained over the baseline features. Therefore, Teager energy-based features do contain additional relevant information for ASR than the magnitude spectrun-based features. The future work could be towards using further sophisticated speech enhancement techniques for far-field conditions and use end-to-end speech recognition systems to improve the performance further by exploiting the Teager energy based features.

ACKNOWLEDGMENTS

The authors would like to thank the Samsung Research Institute, Bangalore (SRI-B) and authorities of DA-IICT Gandhinagar for their kind support to carry out this research work. They also thank University Grants Commission (UGC, New Delhi, India) for providing Rajiv Gandhi National Fellowship (RGNF). The work was done when the first author was research intern at SRI-B and Ph.D student at DA-IICT, Gandhinagar, India.

REFERENCES

[1] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, Orlando, Florida, USA, 2002, pp. IV–4072.

[2] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 127–140, 2012.

[3] S. Dupont, C. Ris, and D. Bachelart, "Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise," in *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, Norwich, UK, 2004.

[4] M. R. Kamble and H. A. Patil, "Analysis of reverberation via teager energy features for replay spoof speech detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 2607–2611.

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA, 2015, pp. 504–511.

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

[7] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech nonlinearities, modulations, and energy operators," in *IEEE ICASSP*, Toronto, Ontario, Canada, 1991, pp. 421–424.

[8] F. Jabloun, A. E. Cetin, and E. Erzin, "Teager energy based feature parameters for speech recognition in car noise," *IEEE Signal Processing Letters*, vol. 6, no. 10, pp. 259–261, 1999.

[9] H. B. Sailor and H. A. Patil, "Auditory feature representation using convolutional restricted boltzmann machine and teager energy operator for speech recognition," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. EL500–EL506, 2017.

[10] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "On separating amplitude from frequency modulations using energy operators," in *IEEE ICASSP*, vol. 2, San Francisco, California, USA, 1992, pp. 1–4.

[11] M. R. Kamble, M. V. S. Krishna, A. K. S. Pulikonda, and H. A. Patil, "Novel teager energy based subband features for audio acoustic scene detection and classification," in *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 2019, pp. 436–444.

[12] M. Kamble, "Design of spoof speech detection system: Teager energy-based approach," Ph.D. dissertation, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India, Feb, 2021.

[13] M. R. Kamble and H. A. Patil, "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.

[14] M. R. Kamble, H. Tak, and H. A. Patil, "Effectiveness of speech demodulation-based features for replay detection." in *INTERSPEECH*, Hyderabad, India, 2018, pp. 641–645.

[15] S. Mallat, *A Wavelet Tour of Signal Proc.* 3^{rd} Edition, Academic press, 1999.

[16] J. Klapper and C. Harris, "On the response and approximation of Gaussian filters," *IEEE IRE Transactions on Audio*, vol. 7, no. 3, pp. 80–87, 1959.

[17] F. Jabloun and A. E. Cetin, "The Teager energy based feature parameters for robust speech recognition in car noise," in *IEEE ICASSP*, vol. 1, 1999, pp. 273–276.

[18] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager energy cepstrum coefficients for robust speech recognition," in *European Conference on Speech Communication and Technology (EUROSPEECH)*, Lisbon, Portugal, 2005, pp. 3013–3016.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE ICASSP*, Brisbane, QLD, Australia, 2015, pp. 5206–5210.

[20] D. Povey *et al.*, "The kaldi speech recognition toolkit," in *IEEE Signal Processing Society workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, HI, USA, 2011.

[21] S. Nayak, D. B. Shashank, S. Bhati, K. Bramhendra, and K. S. R. Murty, "Instantaneous frequency features for noise robust speech recognition," in *National Conference on Communications (NCC)*. IEEE, 2019, pp. 1–5.

[22] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

[23] S. Nayak, S. Bhati, and K. S. R. Murty, "An investigation into instantaneous frequency estimation methods for improved speech recognition features," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP) 2017*, pp. 363–367.

[24] P. Agrawal and S. Ganapathy, "Modulation filter learning using deep variational networks for robust speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 244–253, 2019.

[25] N. Ma, R. Marxer, J. Barker, and G. J. Brown, "Exploiting synchrony spectra and deep neural networks for noise-robust automatic speech recognition," in *IEEE ASRU*, Scottsdale, Arizona, USA, 2015, pp. 490–495.

[26] L. Pfeifenberger, T. Schrank, M. Zohrer, M. Hagmüller, and F. Pernkopf, "Multi-channel speech processing architectures for noise robust speech recognition: 3rd chime challenge results," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA, 2015, pp. 452–459.

[27] D. Bagchi *et al.*, "Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition," in *IEEE ASRU*, Scottsdale, Arizona, USA, 2015, pp. 496–503.

[28] T. Hori *et al.*, "The merl/sri system for the 3rd chime challenge using beamforming, robust feature extraction, and advanced speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA, 2015, pp. 475–481.

[29] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi *et al.*, "The ntt chime-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA, 2015, pp. 436–443.

[30] S. Sivasankaran *et al.*, "Robust ASR using neural network based speech enhancement and feature simulation," in *IEEE ASRU*, Scottsdale, Arizona, USA, 2015, pp. 482–489.