Teager Energy Cepstral Coefficients For Classification of Dysarthric Speech Severity-Level

Aastha Kachhi[1], Anand Therattil[1], Ankur T. Patil[1], Hardik B. Sailor[2], Hemant A. Patil[1]
[1]Speech Research Lab DA-IICT, Gandhinagar
[2]Samsung Research Institute, Banglore, India
Email:{aastha_kachhi, anand_therattil, ankur_patil, hemant_patil}@daiict.ac.in, h.sailor@samsung.com

Abstract—**Dysarthria is a neuro-motor speech impairment that renders speech unintelligibility, which is generally imperceptible to humans *w.r.t* severity-levels. Dysarthric speech classification acts as a diagnostic tool for evaluating the advancement in a patient's severity condition and also aids in automatic dysarthric speech recognition systems (an important assistive speech technology). This study investigates the significance of Teager Energy Cepstral Coefficients (TECC) in dysarthric speech classification using three deep learning architectures, namely, Convolutional Neural Network (CNN), Light-CNN (LCNN), and Residual Networks (ResNet). The performance of TECC is compared with state-of-the-art features, such as Short-Time Fourier Transform (STFT), Mel Frequency Cepstral Coefficients (MFCC), and Linear Frequency Cepstral Coefficients (LFCC). In addition, this study also investigate the effectiveness of cepstral features over the spectral features for this problem. The highest classification accuracy achieved using UA-Speech corpus is 97.18%, 94.63%, and 98.02% (i.e., absolute improvement of 1.98%, 1.41%, and 1.69%) with CNN, LCNN, and ResNet, respectively, as compared to the MFCC. Further, we evaluate feature discriminative capability using $F1$-score, Matthew's Correlation Coefficient (MCC), Jaccard index, and Hamming loss. Finally, analysis of latency period *w.r.t.* state-of-the-art feature sets indicates the potential of TECC for practical deployment of the severity-level classification system.**

Index Terms: Dysarthria, UA-Speech Corpus, TEO Profiles, TECC. [1]

I. Introduction

Coordination between the brain and the speech producing muscles is required for speech production and perception [1]. Speech disorders, such as aparaxia, dysarthria, and stuttering, affect a person's ability to generate speech sounds and formulate intelligible words. These disorders can be caused by neurological or neurodegenerative diseases, such as Cerebral Palsy or Parkinson's disease. It can be mild, moderate, or severe, depending on the impact on the brain. In the case of mild severity, there may be a few minor mispronunciations. On the other hand, in a severe case, there is inability to produce intelligible speech. Among these speech disorders, dysarthria is a relatively common speech disorder [2]. Dysarthria is a neuro-motor speech disability that causes the muscles useful in producing speech to weaken. Additionally, patient's lips, tongue, throat, and upper respiratory tract system are also affected due to brain damage, Cerebral Palsy, muscular dystrophy, or stroke affects, which are linked to dysarthria [3].

Dysarthric severity depends on the damage to the neurological area and its treatment depends on type, underlying cause, severity-level, and symptoms [4]. This motivates researchers to develop diagnostic assistive speech tools for dysarthria intelligibility classification. This problem has been studied in the literature using either speech recognition-based approaches or blind intelligibility assessment. In [5], Mel Frequency Cepstral Coefficients (MFCC) are also employed due to their capacity to capture "global" spectral envelope properties for perceptually-motivated audio classification tasks. In addition, glottal excitation source parameters derived from quasi-periodic sampling of vocal tract systems are also investigated in [6]. As in [7], the disparity in vocal fold vibration between dysarthric and normal speech production cannot be described solely by the rate of vibration (i.e., pitch source information), the *mode* of vibration of the vocal folds are also impacted. Hence, information generated by the waveform of the acoustic speech excitation and glottal flow may contain useful information. Teager Energy Operator (TEO) is known to capture the non-linear excitation source information related to glottal flow waveform of the vocal folds [8], [9]. The key objective of this study is to explore and analyse the difference in non-linearities present in the normal *vs.* dysarthric speech production mechanism using TEO. To that effect, we propose the novel approach in classifying the dysarthric speech severity-level using Teager Energy Cepstral Coefficients (TECC) feature set, which was originally used for speech recognition applications [10], [11]. Many recent studies reveal that the feature representation of the speech signal developed using TEO is useful for anti-spoofing [12], [13].

In this study, two state-of-the-art features are used, namely, MFCC and Linear Frequency Cepstral Coefficients (LFCC) along with comparison with the spectral features mentioned in [14]. The CNN, LCNN, and ResNet are trained using features extracted from speech utterances present in UA-Speech corpus. This study explores the effectiveness of TECC in capturing the non-linearities for dysarthric speech severity-level classification. Speech enhancers are designed for the formants in dysarthric speech enhancement [15] and hence, we present analysis of TEO profile around 1^{st} formant frequency. To the best of the authors' knowledge, this is the first study of its kind, where TECC is proposed for classifying the severity-level of dysarthric speech.

[12]This work was done when Dr. Hardik B. Sailor was at Samsung Research Institute, Bangalore (SRI-B), India in collaboration with DA-IICT, Gandhinagar.

978-616-590-477-3 ©2022 APSIPA

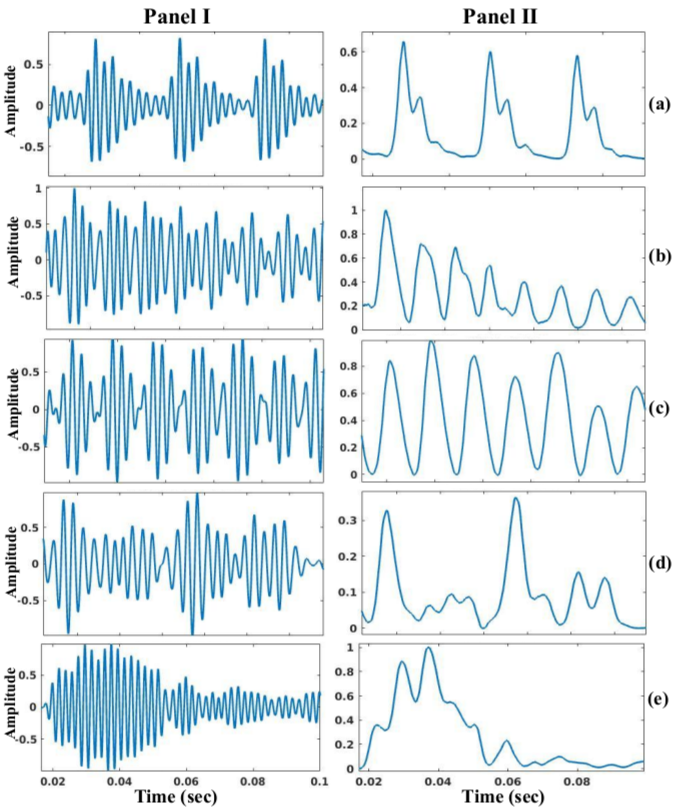

Fig. 1. Subband filtered signal for vowel /i/ by a male speakers around $1^{st} Formant = 500Hz$ (Panel I) and corresponding TEO profile (Panel II) for (a) normal, dysarthic speech with severity-level as (b) very low, (c) low, (d) medium, and (e) high. After [16].

II. Proposed TECC Feature Set

A. Analysis of TEO Profiles

In the signal processing literature, energy of the speech signal $x(t)$ is estimated through Squared Energy of the signal, i.e., the integral of the square of absolute operation over the entire signal under analysis [18]. This method of estimating energy is based on linear filtering theory (in particular, Parseval's energy equivalence), which can describe *only* the linear components of speech production mechanism [16]. However, in particular consider a discrete-time speech signal $x(n)$. The parseval's energy equivalence in Discrete Time Fourier Transform (DTFT) framework is given by [19],

$$\sum_{[n]} |x(n)|^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} |X(e^{j\omega})|^2 \tag{1}$$

$$\sum_{n=-\infty}^{\infty} x(n).x^*(n) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(e^{j\omega}).X^*(e^{j\omega}) \ d\omega \tag{2}$$

From the Equation 2, it can be inferred that

$$< x(n), x(n) >= \frac{1}{2\pi} < X(e^{j\omega}), X(j\omega) > \tag{3}$$

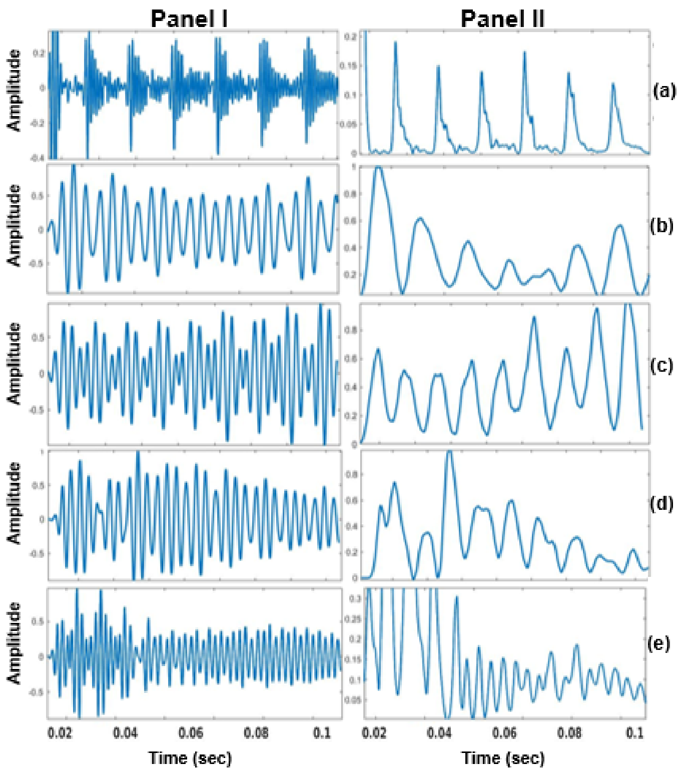

Fig. 2. Subband filtered signal for vowel /e/ by a male speakers around $1^{st} Formant = 500Hz$ (Panel I) and corresponding TEO profile (Panel II) for (a) normal, dysarthic speech with severity-level as (b) very low, (c) low, (d) medium, and (e) high. After [16].

Hence,

$$x(n) * x(\bar{n}) = \frac{1}{2\pi} < X(e^{j\omega}), X(e^{j\omega}) > \tag{4}$$

where $*$ is convolutional operator *w.r.t.* LTI operator, and $< >$ represents the inner product space between two signals. Thus it can be observed that L^2 norm, (i.e.), energy of a signal imposes a linear product structure on the speech signal and this in turn imposes linear structure on the data through convolution operation.

However, because the speech production mechanism is nonlinear, the energy of the speech wave could not be effectively approximated using linear filter theory [20]. TEO was developed to address this problem [21]. It is a nonlinear differential operator that can capture the nonlinear feature of the speech production mechanism as well as the properties of the airflow pattern in the vocal tract system during speech production process [18], [22]. By approximating the derivative operation in continuous-time with backward difference in discrete-time, we obtain the TEO for discrete-time signal $x(n)$ having amplitude, A and monocomponent angular frequency, Ω_m as follows [21]:

$$\Psi[x(n)] = x^2(n) - x(n-1)x(n+1) \approx A^2\Omega_m^2. \tag{5}$$

Where $\Psi[.]$ represents the TEO of monocomponent signal. Furthermore, we analyse the TEO profiles around the 1^{st} formant frequency (i.e., $F1 = 500Hz$) for the utterance of vowel /i/ as show in Figure 1 and vowel /e/ as show in

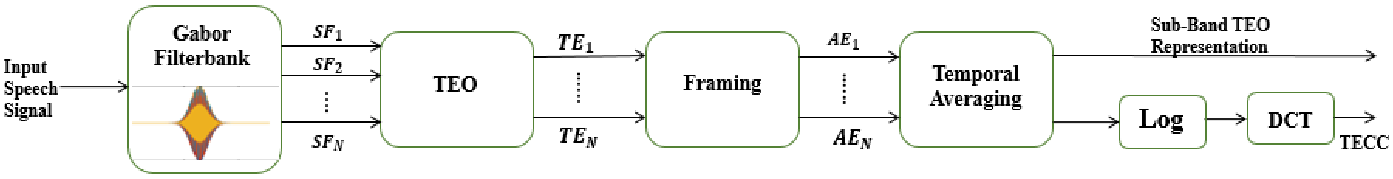

Fig. 3. Functional block diagram of the proposed Subband TEO representation and TECC feature set. (SF: Subband filtered signal, TE: Teager energies, AE: Averaged energies over frames). After [10], [17].

Figure 2 for normal *vs.* different dysarthric severity-levels. Panel-I of Figure 1 and 2 shows the subband filtered signal around 1^{st} formant frequency using a linear-spaced Gabor filter, and Panel-II shows corresponding TEO profiles. Figure 1(a), Figure 1(b), Figure 1(c), Figure 1(d), and Figure 1(e) shows the analysis for normal, very low, low, medium, and high severity-levels, respectively. Similarly, for Figure 2(a), Figure 2(b), Figure 2(c), Figure 2(d), and Figure 2(e) shows the analysis for normal, very low, low, medium, and high severity-levels, respectively. It can be observed from both Figure 1 and Figure 2 that TEO profile for normal speech shows bumps within two consecutive Glottal Closure Instants (GCIs), which are known to indicate non-linearities in speech production mechanism. Furthermore, it can also be observed that the quasi-periodicity in glottal excitation source decreases with increase in severity-level (as observed via aperiodic TEO profile) indicating *disruption* in the rhythmic quasi-periodic movements of vocal folds due to dysarthria. Moreover, it is all the more true in high dysarthric condition. Furthermore, as the severity-level increase, the neuro-motor impairment also increase, which leads to increased vocal fold closure disruption and loosing structural periodicity.

B. TECC Feature Extraction

TEO is derived to find the running estimate of the signal's energy for a monocomponent signal. However, speech signal consists of the frequency range varying from baseband to Nyquist frequencies. Hence, to obtain the monocomponent approximation of the signal, the speech signal is passed through the filterbank, which consists of several subband filters with appropriate center frequency and bandwidth. The subband filtered signals are narrowband signals, which are supposed to approximate the monotone signals and hence, TEO can be applied on these subband filtered signals. In this work, Gabor filterbank with linearly-spaced subbandd filters, is utilized for subband filtering. We chose Gabor filters due to their optimal time and frequency resolution in the framework of Heisenberg's uncertainty principle [16]. TEO is applied on each subband filtered signal to accurately estimate the energy. Furthermore, these narrowband energies are segmented into the frames of 20 ms duration with overlapping of 10 ms. Then, the temporal average for each frame is estimated to produce N-dimensional (D) *subband Teager energy representations (subband-TE)*. Discrete Cosine Transform (DCT) is performed on *subband Teager energy representations* to obtain the TECC. The functional block diagram representation of the proposed subband-TE and TECC feature set is shown in Figure 3. Throughout this study, TECC features extracted using linear scale are termed as TECC.

III. Experimental Setup

A. Dataset Details

The proposed technique is validated using Universal Access dysarthric speech (UA-Speech) Corpus [23]. In our experiments, we have used data of 8 speakers (i.e., 4 males, namely, $M01$, $M05$, $M07$, and $M09$) and 4 females (namely, $F02$, $F03$, $F04$, and $F05$). From 765 word utterances, 465 utterances per speaker as mentioned in [24] was used. For training and testing, we used 90% and 10% of the data, respectively. Table I shows the statistics of UA-Speech Corpus.

TABLE I
Class-wise patient details of UA Speech Corpus. After [23].

	Female	Male
High	F03	M01, M04, M12
Medium	F02	M07, M16
Low	F04	M05, M11
Very Low	F05	M08, M09, M10, M14

B. Feature Sets

In this study, the performance of TECC is compared against MFCC [25], and LFCC [25]. Furthermore, subband-TE being a spectral representation, its performance is compared against the Mel Filterbank (MelFB) coefficients, and Linear Filterbank (LinFB) coefficients. The details of the parameters for these feature sets are given in Table II. All cepstral representations consists of static, Δ, and $\Delta\Delta$ coefficients. In this study, we compare the performance of TECC feature set with the state-of-the-art feature sets, namely, MFCC and LFCC along with its spectral features.

•**MFCC**: The *42*-D MFCC feature set used is extracted using 14 subband filters, which are placed using the Mel scale. Static, Δ, and $\Delta\Delta$ coefficients are considered in this feature set [25].

•**LFCC**: The *120*-D LFCC feature set used is extracted using 40 subband filters, which are placed linearly. Static, Δ, and $\Delta\Delta$ coefficients are considered in this feature set [25]. It is a widely used feature set for speech technology applications. It also mimics the auditory representation, such as Constant-Q

Cepstral Coefficients (CQCC). The windowed speech signal is processed through Fourier transform (FT) to produce STFT. The weighted sum is performed for each Mel scale subband filter. Then, DCT is applied and desired number of cepstral coefficients are extracted to get MFCCs. In this paper, we have used 40 Mel scale subband filters for feature extraction. *13*-D and *40*-D static coefficients have been extracted for two different experiments. However, results with *40*-D static coefficients shows inferior performance compared to that of *13*-D static coefficients. Here, Mel scale filterbank is replaced by linear-scale filterbank, where central frequencies of the subband filters are linearly-spaced. LFCCs are extracted with 40 linear-scale subband filters. All 40 cepstral coefficients are retained and appended with Δ and $\Delta\Delta$ coefficients to form *120*-D LFCC feature vector.

C. Classifiers Details

According to the experiments reported in [5], CNN performs at par *w.r.t* the other deep neural network (DNN)-based classifiers for UA-Speech corpus. Hence, we employed CNN classifier in this study. CNN model was trained using Stochastic Gradient Descent (SGD) algorithm and 3 convolutional layers each with a kernel size 5×5, and 1 Fully-Connected (FC) layer [26]. The input feature is made of uniform size of $D \times 300$, where D is the dimension of the feature vector. Rectified Linear Activation (ReLU) and a max-pool layer are used. Learning rate of 0.001 and cross-entropy loss is selected for loss estimation.

1) Light Convolutional Neural Network (LCNN): LCNN architecture was also implemented, as it is one of the successful architectures for anti-spoofing task [27], [28], [29]. The experiments were performed on the uniform $D \times 300$ features. LCNN architecture uses Max-Feature-Map (MFM) activation operation, for learning with a few parameters [29]. In this study, we utilized seven convolutional layers having MFM activation function followed by two-fully connected layers. The 1^{st} convolutional layer uses the kernel size of 5×5 and stride of 1×1 and the following convolutional layer has a kernel size of 3×3 and stride of 2×2 with learning rate of 0.001. Weights of the LCNN are initialized using Xavier weight initialization technique [30]. ResNets are one of the popular DNN classifiers and introduced to take the advantage of more DNN by integrating the high/mid/low-level features. ResNets are introduced to alleviate the issue of vanishing/exploding gradients of more DNNs. It utilizes the identity mapping as explained in [31], which allows stacking more number of layers without introducing the vanishing/exploding gradients and permits the possibility of smooth convergence. The increase in layers of DNN allow learning high-level features and thus, improving the performance of the system. We have utilized 22 layers ResNet architecture.

TABLE II
DETAILS OF PARAMETERS OF THE VARIOUS FEATURE SETS USED

Parameters	MFCC	LFCC	TECC	MelFB	LinFB	Subband-TE
Frequency Scale	Mel	Linear	Linear	Mel	Linear	Linear
Subband Filter	40	40	40	40	40	40
Feature Dimension	42	120	120	40	40	40

D. Performance Evaluation Metrics

The performance of TECC *w.r.t.* other feature sets are analysed using various statistical parameters, such as $F1$-score, Matthew's Correlation Coefficient (MCC), Jaccard index, and Hamming loss.

1) $F1$-Score: $F1$-score is as widely used measure to the test accuracy of a model, which ranges from 0 to 1, where closer to 1 F1-score indicates better model. It is estimated by taking the harmonic mean of model's precision and recall, as in [33].

2) MCC: MCC is a balanced statistical measure, which measures the effectiveness of the model prediction. It measures the degree of correlation between the actual and predicted class values. MCC ranges between -1 to 1 [34].

3) Jaccard index: It is a metric for determining similarity and dissimilarity between classes. Jaccard index ranges between 0 and 1. The Jaccard index is defined as [35]:

$$\text{Jaccard index} = \frac{TP}{TP + FP + FN}, \tag{6}$$

where TP, FP, and FM represents True Positive, False positive, and False Negative, respectively.

4) Hamming Loss: It takes into account incorrectly predicted class labels. The prediction error (an incorrect label is predicted) and missing error (a relevant label is not predicted) are standardized across the total number of classes and data under test. Hamming Loss can be estimated as [36]:

$$\text{Hamming Loss} = \frac{1}{nL} \sum_{i=1}^{n} \sum_{j=1}^{L} I(y_i^j \neq \hat{y}_i^j), \tag{7}$$

where y_i^j and $\hat{y}_i^j$ are the actual and predicted labels, and I is an indicator function. The more it is close to 0, the better is the performance of the algorithm.

IV. EXPERIMENTAL RESULTS

A. Visualization of Features Space using Linear Discriminant Analysis (LDA)

Severity-level classification capability of the TECC is also validated using the scatter plot obtained using Linear Discriminant Analysis (LDA) due to it's higher image resolution and better projection of the given higher-dimensional feature space to lower-dimensional than the scatter plots obtained using TSNE plots. [32]. Here, we project TECC, MFCC, and LFCC feature sets to 2-D space to obtain the scatter plots for various severity-levels in dysarthria. Figure 4(a), Figure 4(b), and Figure 4(c) shows the scatter plots obtained for MFCC, LFCC, and TECC, respectively, using LDA. It can be observed from Figure 4 that the interclass distance between the clusters of various classes is larger for TECC as compared to the other features. Also, the clusters obtained using TECC are more compact, indicating the better severity-level classification capability of the TECC.

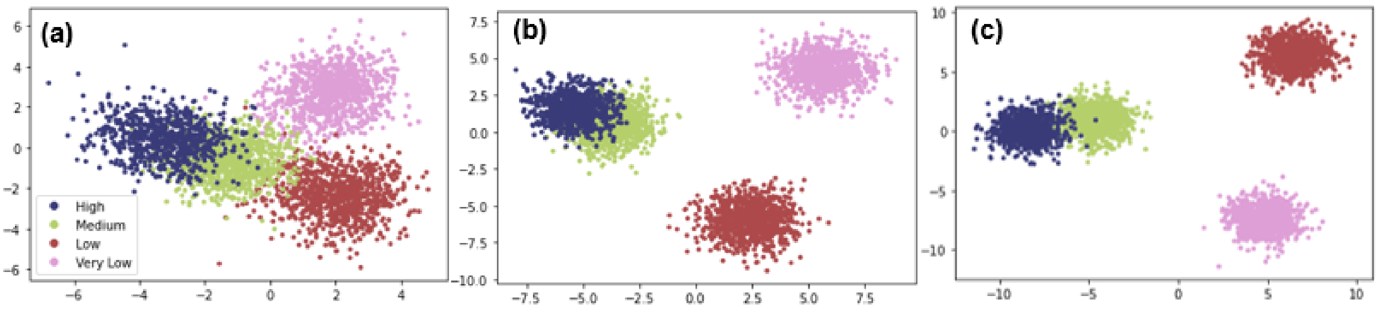

Fig. 4. Scatter plots obtained using LDA for (a) MFCC, (b) LFCC, and (c) TECC. After [32]. Best viewed in colour.

B. Performance Evaluation

The results obtained in % classification accuracy using various features sets and classifiers are reported in Table III. It can be observed that the TECC performs relatively better than the baseline STFT with classification accuracy of 97.12%, 94.63% and 98.02% (i.e., absolute improvement of 5.35 %, 6.20 %, and 2.70 %) for CNN, LCNN, and ResNet classifiers, respectively. Furthermore, it was also observed that there was decrease in % classification accuracy by varying parameters in CNN model. This might be due to overfitting of the model. Furthermore, TECC performs better than baseline STFT features for CNN, LCNN, and ResNet classifiers explored in [24]. Moreover, it was observed that optimum results of TECC were obtained for linear scale. The analysis provided in sub-Section II-A along with experimental results obtained using various classifiers shows that the TECC can be the best possible choice for the severity-level classification of dysarthric speech.

TABLE III
RESULTS (IN % CLASSIFICATION ACCURACY) FOR VARIOUS CLASSIFICATION SYSTEMS. TECC → LINEAR FREQUENCY SCALE USED

Feature Set ↓	**% Classification Accuracy**		
	CNN	**LCNN**	**ResNet**
STFT	91.76	88.43	95.32
MFCC	95.20	93.22	96.33
LFCC	96.32	94.07	97.17
TECC-Mel	92.37	85.87	93.09
TECC	**97.12**	**94.63**	**98.02**
MelFB	96.04	91.24	97.45
LinFB	94.91	89.26	97.17
Subband-TE	95.48	93.22	95.12

As mentioned in [14], the cepstral features perform better on noisy signal. In [23], the noise in dysarthric speech increases with increase in severity-levels. Hence, experiments were also performed on the spectral features *w.r.t* proposed and baseline features with all the three classifiers. It was observed that the cepstral features gave remarkably better % classification accuracy on all the classifiers. Hence, it can be inferred that more the severity-level, more is the speech production noise.

Furthermore, Table IV shows the confusion matrices for the TECC, MFCC, and LFCC for ResNet model. It can be observed that TECC reduces the misclassification errors, especially for high severity-level dysarthria, and overall performance of the TECC is relatively better than the MFCC, and LFCC. Furthermore, $F1$-score, MCC, Jaccard index, and Hamming loss are estimated for all the cepstral features as shown in Table V. It can be observed from Table V that the TECC feature set outperforms the other cepstral features for all the evaluation metrics, indicating relatively better feature discriminative power of TECC.

TABLE IV
CONFUSION MATRIX OBTAINED FOR MFCC, LFCC, AND TECC USING RESNET

Feature	**Severity**	**High**	**Medium**	**Low**	**Very Low**
MFCC	**High**	**72**	0	2	1
	Medium	1	**90**	2	0
	Low	1	1	**88**	3
	Very Low	1	0	0	**92**
LFCC	**High**	**74**	0	1	0
	Medium	1	**88**	2	2
	Low	0	1	**91**	1
	Very Low	1	0	0	**92**
TECC	**High**	**74**	1	0	0
	Medium	1	**92**	0	0
	Low	0	1	**92**	0
	Very Low	1	0	0	**92**

TABLE V
VARIOUS STATISTICAL MEASURES FOR MFCC, LFCC, AND TECC

Feature Sets	**$F1$-Score**	**MCC**	**Jaccard Index**	**Hamming Loss**
MFCC	0.96	0.95	0.93	0.033
LFCC	0.97	0.96	0.95	0.025
TECC	**0.98**	**0.97**	**0.96**	**0.019**

C. Analysis of Latency Period

We analysed latency period for TECC, LFCC, and MFCC feature sets as shown in Figure IV-C. The latency period of the trained model is estimated by computing the % classification accuracy *w.r.t.* varying durations of test speech segment in a test utterance. For analysis of latency period, we chose the duration of the utterances varying from 100 ms to 3000 ms.

To that effect, experiments were performed on $x86\_64$ 32 bit, INTEL(R) Core(TM) i5-2400 CPU at 3.10 GHz. The better performing model w.r.t. latency period should produce the larger accuracy for short speech segments. Moreover, it can be observed that the TECC gave significant % classification accuracy in a limited duration speech utterance of $<$ 500 ms. On the contrary, MFCC and LFCC shows increment in accuracy after a relatively longer utterance duration of 1000 ms. Hence, these results signifies the suitability of TECC for practical dysarthric speech classification system deployment.

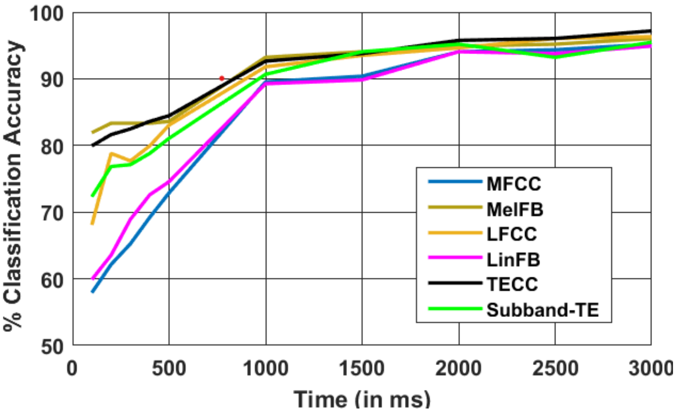

Fig. 5. Latency period *vs.* % classification accuracy comparison between MFCC, MelFB, LFCC, LinFB, TECC, and Subband-TE. Best viewed in colour.

V. Summary and Conclusions

In this study, TECC was proposed to classify the severity-level of dysarthric speech using CNN, LCNN, and ResNet classifiers. To classify the severity-level, TECC captures the non-linearities present in the speech signal. It was observed that the TECC outperforms the other feature sets. This justifies the proposition that as the severity-level increases, the non-linearities decreases and the amount of linear components increases in the speech signal. It can also be seen that due to lack of neuro-motor coordination, the formant structure, which captures the linguistic information are distorted. This study shows that the TEO-based excitation source information are more effective over perceptually-motivated features for this problem. Furthermore, we have also demonstrated the discriminative capability of TECC using statistical measures, such as $F1$-score, MCC, Jaccard index and Hamming loss. For effective analysis of TECC, we analysed the latency period *w.r.t.* state-of-the-art feature sets, which indicates the potential of TECC for practical deployment of severity-level classification system. To the best of authors' knowledge, this study presented the first detailed analysis of TECC for dysarthric speech classification. Our future research efforts will be directed towards evaluation of proposed approach under cross-database scenarios.

VI. Acknowledgements

The authors sincerely thank the PRISM team at Samsung R&D Institute, Bangalore (SRI-B), India. We would also like to thank the authorities of DA-IICT for providing a quality environment and optimum facilities to carry out this work.

References

[1] P. Lieberman, "Primate vocalizations and human linguistic ability," *The Journal of the Acoustical Society of America (JASA)*, vol. 44, no. 6, pp. 1574–1584, 1968.

[2] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assistive Technology*, vol. 22, no. 2, pp. 99–112, 2010.

[3] C. Mackenzie and A. Lowit, "Behavioural intervention effects in dysarthria following stroke: communication effectiveness, intelligibility and dysarthria impact," *International Journal of Language & Communication Disorders*, vol. 42, no. 2, pp. 131–153, 2007.

[4] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of Speech and Hearing Research (JSLHR)*, vol. 12, no. 2, pp. 246–269, 1969.

[5] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification using deep learning frameworks," in 28^{th} *European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands*, 2021, pp. 116–120.

[6] S. Gillespie, Y.-Y. Logan, E. Moore, J. Laures-Gore, S. Russell, and R. Patel, "Cross-database models for the classification of dysarthria presence," in *INTERSPEECH, Stockholm, Sweden*, 2017, pp. 3127–3131.

[7] N. Narendra and P. Alku, "Dysarthric speech classification using glottal features computed from non-words, words and sentences," in *INTERSPEECH, Hyderabad, India*, 2018, pp. 3403–3407.

[8] J. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Albuquerque, NM, USA*, vol. 1, 06 August 2002, pp. 381–384.

[9] L. Gavidia-Ceballos, J. H. Hansen, and J. F. Kaiser, "Vocal fold pathology assessment using AM autocorrelation analysis of the teager energy operator," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, Philadelphia, PA, USA*, vol. 2, 1996, pp. 757–760.

[10] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager energy cepstrum coefficients for robust speech recognition," in *INTERSPEECH*, Lisbon, Portugal, Sept. 2005, pp. 3013–3016.

[11] D. T. Grozdic and S. T. Jovicic, "Whispered speech recognition using deep denoising autoencoder and inverse filtering," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 12, pp. 2313–2322, 2017.

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

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

[14] G. Korvel, O. Kurasova, and B. Kostek, "Comparative analysis of spectral and cepstral feature extraction techniques for phoneme modelling," in *International Conference on Multimedia and Network Information System, Wroclaw, Poland*. Springer, 2018, pp. 480–489.

[15] A. B. Kain, J.-P. Hosom, X. Niu, J. P. Van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 9, pp. 743–759, 2007.

[16] H. M.Teager and S. M.Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *William J. Hardcastle and Alain Marchal (Eds.), Speech Production and Speech Modelling*. Springer, 1990, pp. 241–261.

[17] M. R. Kamble and H. A. Patil, "Detection of replay spoof speech using teager energy feature cues," *Computer Speech & Language*, vol. 65, p. 101140, 2021.

[18] A. V. Oppenheim, A. S. Willsky, S. H. Nawab, G. M. Hernández *et al.*, *Signals & Systems*. Pearson Educación, 1997.

[19] R. Szeliski, *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

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

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

[22] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Pearson Education, 3^{rd} edition, India, 2006.

[23] J. Yu, X. Xie, S. Liu, S. Hu, M. W. Lam, X. Wu, K. H. Wong, X. Liu, and H. Meng, "Development of the CUHK dysarthric speech recognition system for the UA speech corpus." in *INTERSPEECH, Hyderabad, India*, 2018, pp. 2938–2942.

[24] S. Gupta, A. T. Patil, M. Purohit, M. Parmar, M. Patel, H. A. Patil, and R. C. Guido, "Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments," *Neural Networks*, vol. 139, pp. 105–117, 2021.

[25] O. M. Strand and A. Egeberg, "Cepstral mean and variance normalization in the model domain," in *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction, Norwich, United Kingdom*, 30-31 August, 2004.

[26] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of 2010 IEEE Int. Symp. on Circuits and Systems*, Paris, France, 2010, pp. 253–256.

[27] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 82–86.

[28] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC Antispoofing systems for the ASVSpoof2019 challenge," in *INTERSPEECH*, Graz, Austria, Sept. 2019, pp. 1033–1037.

[29] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[30] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA*, 2016, pp. 770–778.

[32] A. J. Izenman, "Linear discriminant analysis," in *Modern Multivariate Statistical Techniques*. Springer, 2013, pp. 237–280.

[33] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[34] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[35] M. Bouchard, A.-L. Jousselme, and P.-E. Doré, "A proof for the positive definiteness of the Jaccard index matrix," *International Journal of Approximate Reasoning*, vol. 54, no. 5, pp. 615–626, 2013.

[36] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier, "Regret analysis for performance metrics in multi-label classification: the case of Hamming and subset zero-one loss," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2010, pp. 280–295.