Significance of Energy Features for Severity Classification of Dysarthria

Chapter *in* Lecture Notes in Computer Science · November 2022

DOI: 10.1007/978-3-031-20980-2_28

5 authors, including:

Aastha Kachhi
HCL
15 PUBLICATIONS **48** CITATIONS
SEE PROFILE

Anand Therattil
Govivace
6 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Ankur Patil
Dhirubhai Ambani Institute of Information and Communication Technology
23 PUBLICATIONS **228** CITATIONS
SEE PROFILE

Hemant Patil
Dhirubhai Ambani Institute of Information and Communication Technology
303 PUBLICATIONS **3,361** CITATIONS
SEE PROFILE

Significance of Energy Features for Severity Classification of Dysarthria

Aastha Kachhi[1](✉), Anand Therattil[1], Ankur T. Patil[1], Hardik B. Sailor[2], and Hemant A. Patil[1]

[1] Speech Research Lab, DAIICT, Gandhinagar, India
{aastha_kachhi,anand_therattil,ankur_patil,hemant_patil}@daiict.ac.in
[2] Samsung R&D Bangalore, Bengaluru, India

Abstract. Dysarthria is a neuro-motor speech disorder that affects the intelligibility of speech, which is often imperceptible depending on its severity-level. Patients' advancement in the dysarthric severity-level are diagnosed using the classification system, which also aids in automatic dysarthric speech recognition (an important assistive speech technology). This study investigates presence of the linear *vs.* non-linear components in the dysarthic speech for severity-level classification using the Squared Energy Cepstral Coefficients (SECC) and Teager Energy Cepstral Coefficients (TECC), which captures the linear and non-linear production features of the speech signal, respectively. The comparison of the TECC and SECC is presented *w.r.t* the baseline STFT and MFCC features using three deep learning architectures, namely, Convolutional Neural Network (CNN), Light-CNN (LCNN), and Residual Neural Network (ResNet) as pattern classifiers. SECC gave improved classification accuracy by 6.28% (7.89%/3.60%) than baseline STFT system, 1.7% (4.23%/0.99%) than MFCC and 0.1.41% (0.56%/0.28%) than TECC on CNN (LCNN/ResNet) classifier systems, respectively. Finally, the analysis of feature discrimination power is presented using Linear Discriminant Analysis (LDA), Jaccard index, Matthew's Correlation Coefficient (MCC), $F1$-score, and Hamming loss followed by analysis of latency period in order to investigate practical significance of proposed approach.

Keywords: Dysarthria · UA-Speech corpus · TEO profiles · TECC · Squared energy · SEO profiles

1 Introduction

Speech production requires proper coordination between speech producing muscles and brain [17]. Speech disorders, such as aparaxia, dysarthria, and stuttering are caused by a lack of this coordination. These conditions affect a person's capacity to produce speech sounds. These disorders are categorized as neurological or

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

S. R. M. Prasanna et al. (Eds.): SPECOM 2022, LNAI 13721, pp. 325–337, 2022.
https://doi.org/10.1007/978-3-031-20980-2_28

neurodegenerative disorders. These disorders might be mild or severe, depending on the impact and damage done on the area of brain. Amongst these disorders, dysarthria is the most common disorder [28]. Dysarthria is a neuro-motor disorder, which weakens the muscles that are responsible for speech production. Apart from the muscles, articulators such as lips, tongue, throat, and upper respiratory tract system of a patient are also affected. Severity of dysarthria depends on intensity of damage done to neurological area, and its treatment also depends on the cause and symptoms [2]. These factors have inspired researchers to create a diagnostic aid for the enhancement of speech intelligibility for dysarthic speech.

Extensive study for this problem has been explored in the literature. These studies employed state-of-the-art Mel Frequency Cepstral Coefficients (MFCC) because of their ability to capture *global* spectral envelope for a perceptually-motivated audio classification tasks [11]. Moreover, studies based on glottal source parameters derived from quasi-periodic sampling of vocal tract systems are also explored in [6]. The mismatch in vocal fold vibration between dysarthric and normal speech production, as indicated in [20], cannot be explained merely by the rate of vibration (i.e., excitation source information); rather the *mode* of vibration (oscillations) of vocal folds is also influenced. As a result, information generated by the waveform of acoustic speech excitation and glottal flow may have useful information. Moreover, there is difference in non-linearities in speech production for normal *vs.* dysarthric speech. Hence, the speech signal energies could not be estimated accurately using linear filter theory [26]. Hence, to capture the non-linearities present during the speech production mechanism, Teager Energy Operator (TEO) was proposed in 1990 [19]. Many recent studies reveal that feature representation using TEO are useful for anti-spoofing task [4]. To that effect, the key objective of this study is to explore and analyse the difference in non-linearities present in normal and various severity-level dysarthric speech using Teager Energy Cepstral Coefficients (TECC) [4,7], and Squared Energy Cepstral Coefficients (SECC).

To validate this hypothesis, we compare the performance of SECC with TECC. The Short-Time Fourier Transform (STFT) is used as baseline for this study as in [11]. The features extracted from the word utterances spoken by speakers from standard UA-Speech corpus were used to train CNN, LCNN, and ResNet models. Dysarthric speech is enhanced using speech enhancer designed for formants [12] and hence, this study also investigates the proposition by analysing the TEO profile and Squared Energy Operator (SEO) around 1^{st} formant frequency for vowel, $/i/$, and $/e/$.

The rest of the paper is organized as follows: Sect. 2 presents the TEO *vs.* SEO analysis alaong with feature extraction process whereas Sect. 3 gives the Experimental setup. Section 4 gives the detailed experimental results and analysis. Finally Sect. 5 concludes the paper along with potential future direction.

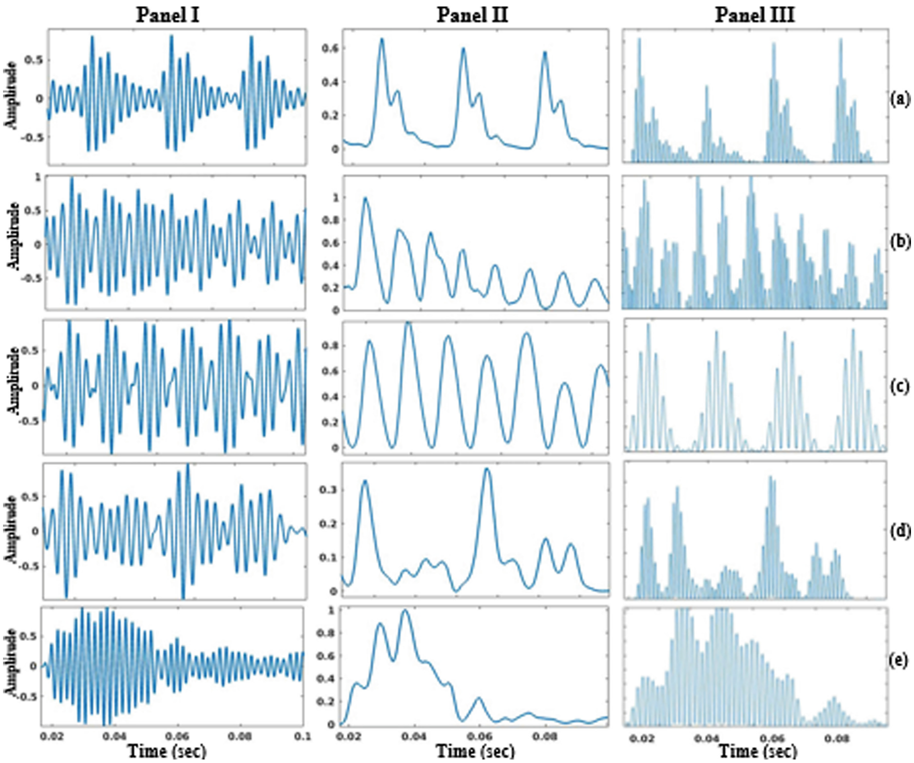

Fig. 1. Subband filtered signal (for vowel /i/) for male speakers around $1^{st} formant F_1 = 500$ Hz (Panel I), corresponding TEO profile (Panel II), and corresponding $|.|^2$ (i.e., SEO profile) envelope (Panel III) for (a) normal, dysarthic speech with severity-level as (b) very low, (c) low, (d) medium, and (e) high. After [19].

2 TEO *vs.* SEO

2.1 Analysis of SEO and TEO Profile

In the signal processing literature, the energy of the speech signal $x(t)$ is estimated by calculating the integral of square of absolute operation across the entire signal under consideration, i.e., estimating the squared energy of the signal, referred to as SEO [21]. This energy estimation method is based on linear filtering theory (specifically, Parseval's energy equivalence, the total energy of a signal, i.e., squared energy is conserved in the frequency-domain and this is also the condition of existence of inverse for several *linear* transforms, such as Fourier, Gabor, and Wavelet transforms), which can only represent the linear components of the speech generation process [19]. 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 [25]:

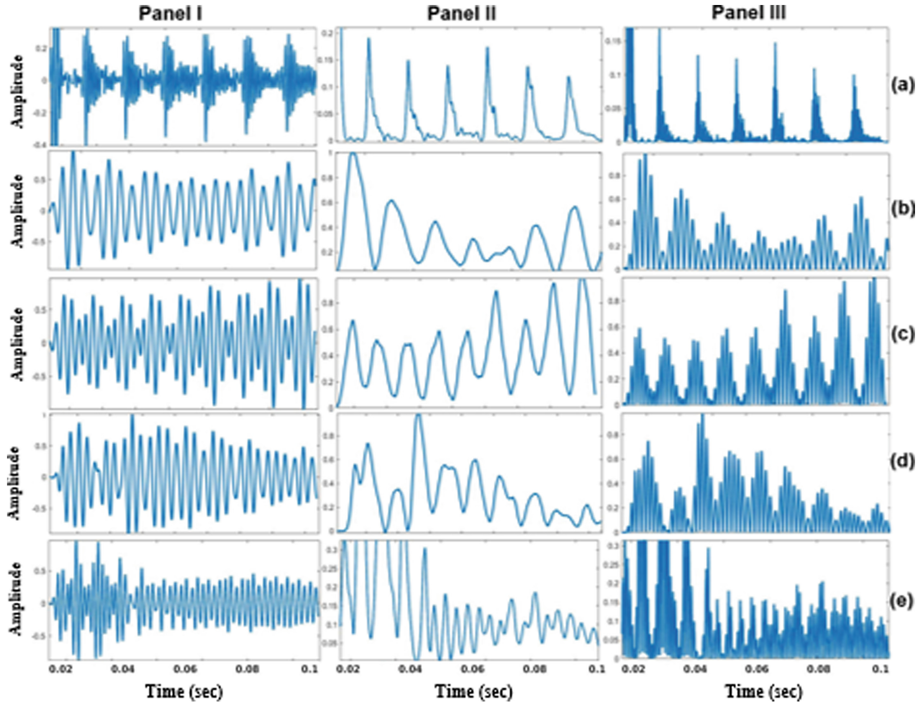

Fig. 2. Subband filtered signal (for vowel /e/) for male speakers around 1^{st} formant $F_1 = 500$ Hz (Panel I), corresponding TEO profile (Panel II), and corresponding $|.|^2$ (i.e., SEO profile) envelope (Panel III) for (a) normal, dysarthic speech with severity-level as (b) very low, (c) low, (d) medium, and (e) high. After [19].

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

where $*$ denotes the complex conjugate operation, and $X(e^{j\omega})$ represents DTFT of $x(n)$. From the Eq. (1), it can be inferred that

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

where $<,>$ indicates inner product operator. Equation 2 can be represented as

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

where $\bar{x}(n) = x^*(-n)$, and in Eq. (3) $*$ is convolutional operator *w.r.t.* LTI operator. Thus it can be observed that the square of the L^2 norm, (i.e.), energy of a signal imposes an inner product structure on the speech signal and this in turn imposes linear structure on the data through convolution operation.

The energy of the speech wave could not be properly approximated using linear filter theory because the true speech production mechanism is non-linear [26]. Hence, TEO was developed to alleviate this issue [13]. It is defined as a nonlinear differential operator that can represent the speech production mechanism as well as the features of the airflow pattern in the vocal tract system during speech production [21,23]. The TEO for a discrete-time signal $x(n)$ with amplitude A and monocomponent angular frequency ω is obtained by approximating the derivative operation in continuous-time with backward difference in discrete-time [13]. In particular,

$$\Psi\{x(n)\} = x^2(n) - x(n-1)x(n+1) \approx A^2\Omega_m^2. \tag{4}$$

Here, we analyse the TEO profiles around the 1^{st} formant frequency (i.e., $F_1 = 500$ Hz) for the utterance /*i*/ and /*e*/ for normal *vs.* different dysarthric severity-levels. Panel I of Fig. 1 and Fig. 2 shows the subband filtered signal around 1^{st} formant (F_1) frequency using a linearly-spaced Gabor filter, and Panel II shows corresponding TEO profiles. Figure 1 and Fig. 2(a), Fig. 1 and Fig. 2(b), Fig. 1 and Fig. 2(c), Fig. 1 and Fig. 2(d), and Fig. 1 and Fig. 2(e) shows the analysis for normal, very low, low, medium, and high severity-levels, respectively. The high energy pulses shown in the TEO profiles are around GCIs. The *sudden* closure of glottis provides the impulse-like excitation to vocal tract. TEO, by virtue of it's property, captures the high energy strength due to sudden closure of glottis and airflow through the glottis. The region around the sudden closure point also has relatively high energy [22]. As observed from Fig. 1(a) and Fig. 2(a) the TEO profiles shows the non decaying function with *bumps* within glottal cycle. Therefore, the presence of bumps within the glottal cycle indicates that the speech production process is not linear only due to the linear model, but also includes a significant contribution from nonlinear effects (captured through the TEO profile, which may not be well captured through the linear model alone). We refer to this contribution as the aeroacoustic contribution [13,19,23]. It can be observed that TEO profile for normal speech shows *bumps* within two consecutive Glottal Closure Instants (GCIs), which are known to indicate non-linearities in the speech production mechanism [23]. 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 significant in high severity dysarthric condition. Furthermore, as the severity-level increases, the neuro-motor impairment also increase, which leads to increased disruption in vocal fold closure and loosing *structural* periodicity. From Panel III of Fig. 1 and Fig. 2, which shows the SEO profiles around 1^{st} formant frequency for vowel /*i*/ and /*e*/ respectively, it can be observed that the SEO is capable of maintaining the periodicity in the speech produced by dysarthric speaker, which are not captured by TEO due to possible decrease in non-linearities. Hence, it can be said that as the dysarthric severity-level increases, the linearities (i.e., linear component) in speech signal increases.

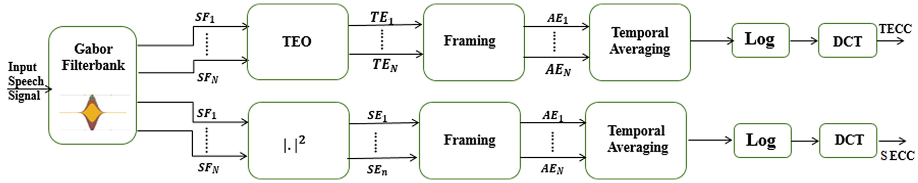

Fig. 3. Functional block diagram of the proposed TECC and SECC feature sets. (SF: Subband filtered signal, SE: Squared linear energies, TE: Teager energies, AE: Averaged energies over frames). After [4].

2.2 SECC and TECC Feature Extraction

TEO was originally derived to find the running estimate of the signal's energy for a monocomponent signal [13]. However, speech signal consists of the frequency range varying from low frequency to the Nyquist frequency. 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 subband filters are utilized for subband filtering. We chose Gabor filters due to their *optimal* time and frequency resolution in the context of Heisenberg's uncertainty principle in signal processing framework [19]. TEO is applied on each subband filtered signal to accurately estimate the signal's energy. Furthermore, these narrowband energies are segmented into the frames of 20 ms duration with overlapping of 10 ms. Then, temporal average for each frame is estimated to produce N-dimensional (D feature vector) *subband Teager energy representations (subband-TE)*. Discrete Cosine Transform (DCT) is applied on *subband Teager energy representations* to obtain the TECC feature vector.

For SECC extraction, these narrowband output signals from Gabor filterbank are squared to estimate corresponding energies. Next, these narrowband energies are segmented with similar number of frames and window overlap. Temporal averaging for each frame is estimated (i.e., squared energy of each subband signal) to get N-D *subband Squared Energy representation (subband-SE)*. DCT is applied on *subband Squared energy representations* to obtain the SECC. The functional block diagram representation of TECC and SECC feature vectors are shown in Fig. 3. Throughout this paper, TECC and SECC features extracted using linear frequency scale and for both the feature vectors, DCT does the job of feature decorrelation, energy compaction, and dimensionality reduction.

3 Experimental Setup

3.1 Dataset Used

The proposed technique is validated using standard Universal Access dysarthric speech (UA-Speech) Corpus [29]. Table 1 shows the statistics of UA Speech

Corpus. 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 [8] was used. For training and testing, we used 90% and 10% of the data, respectively.

Table 1. Class-wise patient details. After [29].

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

3.2 Details of Feature Sets

In this study, performance of SECC is analysed against three baseline systems, namely, STFT, TECC, and MFCC [8,24]. The STFT features are extracted using a 2 ms window length and 0.5 ms TECC feature set is extracted as mentioned in Subsect. 2.2. The DCT applied on *subband-TE* gives 120-D TECC feature vector which, consists of 40-D static, Δ, and $\Delta\Delta$ coefficients, each. MFCC feature set is extracted by applying STFT on the speech signal. The weighted sum is performed on each Mel scale subband filtered signals and as a result, we get Mel filterbank coefficients. Thereafter, log and DCT are applied to obtain MFCC feature vector of 120-D including 40-D of each, static, Δ, and $\Delta\Delta$ coefficients.

3.3 Classifier Details

Convolutional Neural Network (CNN). Based on the experiments reported in [11], CNN performs comparable *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 Adam optimizer algorithm and 3 convolutional layers each with kernel size 5 × 5, and 1 Fully-Connected (FC) layer [16]. 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.

Light-CNN. The LCNN architecture was also implemented, since it is one of the most successful designs for anti-spoofing tasks [14,15,27]. The experiments were performed on the uniform $D \times 300$ features. LCNN architecture uses Max-Feature-Map (MFM) activation operation, for learning with a small number of parameters [27].

Residual Neural Network (ResNet). 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 [9], 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 classification performance of the system. We have utilized ResNet architecture having 22 layers.

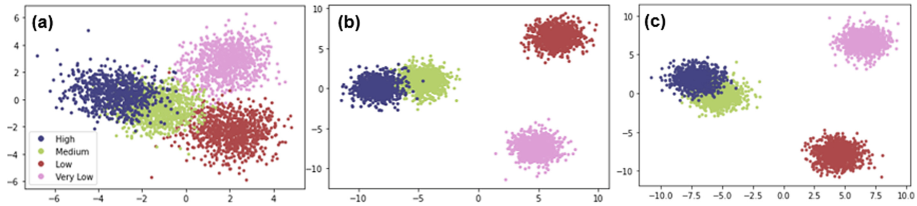

Fig. 4. Scatter plots obtained using LDA for (a) MFCC, (b) TECC, and (c) SECC. After [10]. Best viewed in colour. (Color figure online)

3.4 Performance Evaluation

Performance of SECC is compared *w.r.t.* TECC and MFCC using various statistical measures, such as $F1$-Score, Matthew's Correlation Coefficient (MCC), Jaccard's Index, and Hamming Loss.

F1-Score. It is an extensively used statistical measure to analyse the performance of the model. It is the harmonic mean of the precision and recall of the model, as in [5]. It ranges between 0 and 1, where model with score closer to 1 indicated better performance.

MCC. It gives the correlation degree between predicted and actual class [18]. It is generally considered balanced measure for model comparison. It ranges between −1 and 1.

Jaccard Index. Jaccard Index is a parameter for calculating similarity and dissimilarity between given classes. It ranges between 0 and 1. It is defined as [1]:

$$\text{Jaccard Index} = \frac{TP}{TP + FP + FN}, \tag{5}$$

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

Hamming Loss. It takes into account class labels that were incorrectly predicted. The prediction error (i.e., prediction of an inaccurate label) and missing error (prediction of a relevant label) are normalised over the total number of classes and test data. Hamming loss can be calculated using the following formula [3]:

$$\text{Hamming Loss} = \frac{1}{nL} \sum_{i=1}^{n} \sum_{j=1}^{L} I(y_i^j \neq \hat{y}_i^j), \tag{6}$$

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.

4 Experimental Results

4.1 Performance Evaluation

The results obtained as % classification accuracy using various feature sets are reported in Table 2. It can be observed that SECC performs absolute relative performance than the baseline STFT, MFCC, and TECC with classification accuracy of 6.28% (7.89%/3.6%) than baseline STFT, 1.7% (4.23%/0.99%) than MFCC and 0.1.41% (0.56%/0.28%) than TECC on CNN (LCNN/ResNet) classifier systems, respectively. Furthermore, SECC performs better than the baseline MFCC explored in [11]. The analysis in the subsequent Section, along with the classification accuracy obtained using various classifiers, indicate that the linearities in speech production mechanism increases with increase in dysarthric severity-level.

Furthermore, Table 3 shows the confusion matrices for MFCC, TECC, and SECC for ResNet model. It can be observed that SECC reduces the misclassification errors corresponding to the different severity-levels, indicating the better performance of SECC *w.r.t.* TECC and MFCC. Furthermore, performance of SECC *w.r.t.* TECC and MFCC is also analysed using $F1$-Score, MCC, Jaccard Index, and Hamming Loss as shown in Table 4. It can be observed from Table 4 that SECC performs better than the TECC for the dysarthic severity-level classification.

Table 2. Confusion matrix obtained for STFT, MFCC, TECC, and SECC.

Feature set	% Classification accuracy		
	CNN	LCNN	ResNet
STFT	91.76	88.43	95.32
MFCC	96.32	92.09	95.33
TECC	96.61	95.76	96.04
SECC	**98.02**	**96.32**	**98.92**

Table 3. Confusion matrix for MFCC, TECC, and SECC using CNN.

Feature	Severity	High	Medium	Low	Very Low
MFCC	**High**	**67**	4	3	1
	Medium	2	**90**	0	0
	Low	1	1	**91**	0
	Very Low	1	0	0	**92**
TECC	**High**	**72**	1	2	0
	Medium	2	**90**	0	0
	Low	1	1	**91**	0
	Very Low	0	0	0	**93**
SECC	**High**	**74**	1	0	0
	Medium	2	**90**	0	0
	Low	1	0	**92**	0
	Very Low	0	0	0	**93**

Table 4. Various statistical measures of MFCC, TECC, and SECC.

Feature set	F1-score	MCC	Jaccard index	Hamming loss
MFCC	0.96	0.95	0.82	0.036
TECC	0.97	0.96	0.95	0.025
SECC	**0.98**	**0.97**	**0.96**	**0.019**

4.2 Visualization of Various Features Using Linear Discriminant Analysis (LDA)

Capability of SECC to classify severity-level is also validated by LDA scatter plots 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 t-sne plots [10]. Here MFCC, TECC, and SECC features are projected to the 2-D space to get the scatter plots for various severity-levels of dysarthria. Figure 4(a), Fig. 4(b), and Fig. 4(c) shows the LDA plots of MFCC, TECC, and SECC, respectively. From the Fig. 4, it can be observed that for SECC, the variance of each severity-level clusters is less resulting in relatively better performance of SECC, which increases the interclass distance between the clusters than the baselines STFT, MFCC, and TECC.

4.3 Latency Analysis

Latency period for SECC *w.r.t.* TECC and MFCC were also analysed as shown in Fig. 5. Latency period was calculated using the % classification accuracy on varying test utterance. The utterance was varied from 50 ms to 300 ms. To that

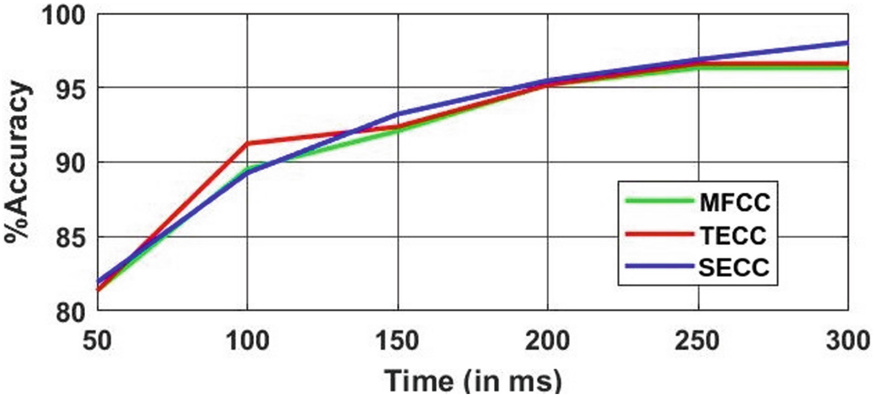

Fig. 5. Latency period *vs.* % classification accuracy comparison between MFCC, TECC, and SECC. Best viewed in colour. (Color figure online)

effect, experiments were performed on x86_64 32 bit, INTEL(R) Core(TM) i5-2400 CPU at 3.10 GHz. For short speech segments, the better performing model should produce higher accuracy in terms of relatively lower latency period. From the Fig. 5, it can be observed that SECC gives consistent and relatively better classification accuracy in short duration of time as 60 ms. Furthermore, TECC and MFCC gives increased classification accuracy for speech segment of 100 ms and 250 ms, respectively. Hence, these results signifies the practical suitability of SECC in dysarthric speech severity-level classification.

5 Summary and Conclusion

In this study, we analysed the effect of the linear *vs.* non-linear energy operator for the analysis and classification of the severity-level of the dysarthric speech. The squared energy operator is analysed against TEO to validate the effect of non-linearity w.r.t. severity-level in dysarthric speech. The bumps which characterizes the non-linearities in the speech production mechanism were observed in the TEO profile of normal speech signal. Whereas, this bumpy structure was found to be reduced w.r.t. increase in the severity-level. Hence, we believe that the squared energy operator seems to be more suitable for dysarthric speech analysis as SEO is known to impose linear structure on the speech data and have help in capturing linear components in speech production. These hypotheses are tested by performing the experiments using CNN, LCNN, and ResNet classifiers. The experimental results showed that the squared energy operator is more suitable over nonlinear operator, such as TEO for dysarthric speech analysis and classification. The observation were validated using various statistical measures, such as $F1$-score, MCC, Jaccard Index, Hamming Loss, LDA, and analysis of latency period. Our future efforts will be directed to extend and validate this work on the other dysarthic speech corpora, such as TORGO, and Home service.

Acknowledgments. 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 Gandhinagar for kind support and cooperation to carry out this research work.

References

1. Bouchard, M., Jousselme, A.L., Doré, P.E.: A proof for the positive definiteness of the Jaccard index matrix. Int. J. Approx. Reason. **54**(5), 615–626 (2013)
2. Darley, F.L., Aronson, A.E., Brown, J.R.: Differential diagnostic patterns of dysarthria. J. Speech Hear. Res. (JSLHR) **12**(2), 246–269 (1969)
3. Dembczyński, K., Waegeman, W., Cheng, W., Hüllermeier, E.: Regret analysis for performance metrics in multi-label classification: the case of hamming and subset zero-one loss. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS (LNAI), vol. 6321, pp. 280–295. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15880-3_24
4. Dimitriadis, D., Maragos, P., Potamianos, A.: Auditory Teager energy cepstrum coefficients for robust speech recognition. In: INTERSPEECH, Lisbon, Portugal, pp. 3013–3016, September 2005
5. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. **27**(8), 861–874 (2006)
6. Gillespie, S., Logan, Y.Y., Moore, E., Laures-Gore, J., Russell, S., Patel, R.: Cross-database models for the classification of dysarthria presence. In: INTERSPEECH, Stockholm, Sweden, pp. 3127–31 (2017)
7. Grozdic, D.T., Jovicic, S.T.: Whispered speech recognition using deep denoising autoencoder and inverse filtering. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) **25**(12), 2313–2322 (2017)
8. Gupta et al., S.: Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. Neural Netw. **139**, 105–117 (2021)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), LV, Nevada, USA, pp. 770–778 (2016)
10. Izenman, A.J.: Linear discriminant analysis. In: Izenman, A.J. (ed.) Modern Multivariate Statistical Techniques, pp. 237–280. Springer Texts in Statistics. Springer, New York (2013). https://doi.org/10.1007/978-0-387-78189-1_8
11. Joshy, A.A., Rajan, R.: Automated dysarthria severity classification using deep learning frameworks. In: 28th European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands, pp. 116–120 (2021)
12. Kain, A.B., Hosom, J.P., Niu, X., Van Santen, J.P., Fried-Oken, M., Staehely, J.: Improving the intelligibility of dysarthric speech. Speech Commun. **49**(9), 743–759 (2007)
13. Kaiser, J.F.: On a simple algorithm to calculate the energy of a signal. In: International Conference on Acoustics. Speech, and Signal Processing (ICASSP), New Mexico, USA, pp. 381–384 (1990)
14. Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., Shchemelinin, V.: Audio replay attack detection with deep learning frameworks. In: INTERSPEECH, Stockholm, Sweden, pp. 82–86, August 2017
15. Lavrentyeva, G., Novoselov, S., Tseren, A., Volkova, M., Gorlanov, A., Kozlov, A.: STC Antispoofing systems for the ASVSpoof2019 challenge. In: INTERSPEECH, Graz, Austria, pp. 1033–37, September 2019

16. LeCun, Y., Kavukcuoglu, K., Farabet, C.: Convolutional networks and applications in vision. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems, Paris, France, pp. 253–256 (2010)
17. Lieberman, P.: Primate vocalizations and human linguistic ability. J. Acoust. Soc. Am. (JASA) **44**(6), 1574–1584 (1968)
18. Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) Protein Struct. **405**(2), 442–451 (1975)
19. Teager, H.M., Teager, S.M.: Evidence for nonlinear sound production mechanisms in the vocal tract. In: Hardcastle, W.J., Marchal, A. (eds.) Speech Production and Speech Modelling. NATO ASI Series, vol. 55, pp. 241–261. Springer, Dordrecht (1990). https://doi.org/10.1007/978-94-009-2037-8_10
20. Narendra, N., Alku, P.: Dysarthric speech classification using glottal features computed from non-words, words, and sentences. In: INTERSPEECH, Hyderabad, India, pp. 3403–3407 (2018)
21. Oppenheim, A.V., Willsky, A.S., Nawab, S.H., Hernández, G.M., et al.: Signals & Systems, 1st edn. Pearson Educación (1997)
22. Patil, H.A., Parhi, K.K.: Development of TEO phase for speaker recognition. In: 2010 International Conference on Signal Processing and Communications (SPCOM), pp. 1–5. IEEE (2010)
23. Quatieri, T.F.: Discrete-Time Speech Signal Processing: Principles and Practice, 3rd edn. Pearson Education, India (2006)
24. Strand, O.M., Egeberg, A.: 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, pp. 30–31, August 2004
25. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer, London (2010). https://doi.org/10.1007/978-1-84882-935-0
26. Teager, H.M.: Some observations on oral air flow during phonation. IEEE Trans. Acoust. Speech Signal Process. **28**(5), 599–601 (1980)
27. Wu, X., He, R., Sun, Z., Tan, T.: A light CNN for deep face representation with noisy labels. IEEE Trans. Inf. Forensics Secur. **13**(11), 2884–2896 (2018)
28. Young, V., Mihailidis, A.: Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: a literature review. Assist. Technol. **22**(2), 99–112 (2010)
29. Yu, J., et al.: Development of the CUHK dysarthric speech recognition system for the UA speech corpus, In: INTERSPEECH, Hyderabad, India, pp. 2938–2942 (2018)