Cross-Teager Energy Cepstral Coefficients For Dysarthric Severity-Level Classification

Anand Therattil, Aastha Kachhi, Hemant A. Patil

Speech Research Lab DA-IICT, Gandhinagar, India

{anand_therattil, aastha_kachhi, hemant_patil}@daiict.ac.in

Abstract

Dysarthria is a degenerative motor speech impairment, generally resulting into neurological damage in human body. This impairment causes the speech to be unintelligible to the humans, depending on the patient's severity-level. Classification of dysarthric severity-level aids as a diagnostic tool to assess advancement of the patient's condition, which also aids in dysarthric Automatic Speech Recognition (ASR), as the traditional ASR systems performs poorly on dysarthric speech. This study investigates the effect of Cross-Teager Energy Cepstral Coefficients (CTECC) on standard and statically meaningful UA-Speech corpus, which captures the energy-based signal from microphone array using the deep learning architecture, such as Convolutional Neural Network (CNN) with classification accuracy of 95.76%. The key objective of this thesis is to select optimal microphone (channel) with minimum amount of energy, which captures the maximum linguistic information of dysarthric speech. Additionally, the performance of CTECC feature is compared with Short-Time Fourier Transform (STFT)-based features, which gave classification accuracy of 91.76% on CNN classifier. Further, the Jaccard index, Matthew's Correlation Coefficient (MCC), $F1$-score, and Hamming loss are used to examine feature discrimination power. Finally, we analyze the latency period for the proposed CTECC feature set for practical deployment of the classification system.
Index Terms: Dysarthria, UA-Speech Corpus, Cross-TEO, CNN.

1. Introduction

For the production of speech sounds, proper coordination between the brain and the speech generating muscles is essential [1]. Speech disorders, such as aparaxia, dysarthria, and stuttering result from a lack of this coordination. These conditions impair a person's ability to produce speaking sounds. Cerebral palsy and Parkinson's disease, for example, are classified as neurological or neuro-degenerative diseases. Depending on the influence on the brain area, the severity of these disorders might range from minor to severe. In a mild case, the patient may misspell a few words, whereas in a severe case, the patient is unable to create understandable speech. Dysarthria is a somewhat prevalent speech issue among these, according to [2]. Dysarthria is a neurological condition that affects speech. People with this condition have weak muscles that create speech. Due to brain injury, dynamic motions of articulators, such as the lips, tongue, throat, and upper respiratory tract system are also impacted. Cerebral palsy, muscular dystrophy, and stroke are some of the other reasons that can induce dysarthria, according to [3].

The impact and damage to the area of neurological injury, which is identified by a brain and nerve test, determines the severity of dysarthria. The type, underlying cause, severity-level, and symptoms all have an impact on the treatment [4]. Researchers are eager to develop supportive methods for dysarthric intelligibility categorization because of the ambiguity in treatment.

Short-Time Fourier Transform (STFT) [5] and numerous acoustical features have been extensively used in the literature to classify dysarthria severity-levels [6]. Due to its ability to capture global spectral envelope features, state-of-the-art feature sets, such as Mel Frequency Cepstral Coefficients (MFCC), were used in [7]. Glottal excitation source parameters obtained from quasi-periodic sampling of the vocal tract system were implemented in [8], in addition to perceptually justified state-of-the-art feature set. Speech signals are considered non-stationary signals in signal processing because of the large and dynamic range of numerous frequency components in short-time spectra. The frequency spectrum changes instantly due to dynamic articulator motions.

The Cross-Teager Energy Operator (CTEO) is an extension of the Teager Energy Operator (TEO), which was developed to determine non-linearities in the speech production mechanism as well as the characteristics of airflow pattern in the vocal tract system. CTEO captures the relative speech production energies between multiple microphone channels. Further, through the CTEO-based Cross-Teager Energy Cepstral Coefficients (CTECC) the optimal channel selection is possible, which can aid in designing efficient, and robust Automatic Speech Recognition (ASR) systems and speech enhancement systems for dysarthric speech [9, 10]. In this work, we illustrate the effect of selecting the channels with maximum and minimum energy through CTECC for dysarthric severity-level classification. Here, different microphone arrays are considered for determining the discriminative cues for the severity-level classification of dysarthria. It is been observed in the literature [11] that the CTECC has better capacity of capturing the linguist information. Hence, this study aims to capture better linguist information for dysarthric severity-level classification using CTECC. To the best of author's knowledge, CTECC has been proposed for the first time for dysarthric severity-level classification.

The rest of the paper is organized as follows: Section 2 presents technical details of proposed CTECC feature set, Section 3 gives the details of experimental setup, whereas Section 4 presents analysis of results. Finally, Section 5 concludes the paper along with future research directions.

2. Proposed CTECC Feature Set

TEO is known to track the instantaneous energy of the speech signal more accurately than the conventional squared energy operator in the signal processing literature (i.e., L^2 norm of the signal). This may be due to the fact that TEO is able to capture the non-linearities in the speech signal [13]. For a monocompo-

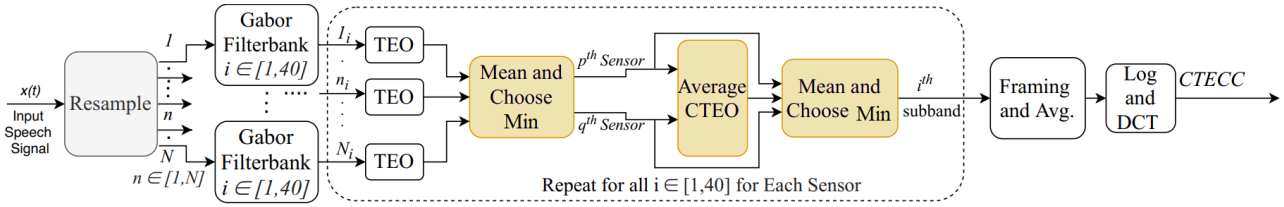

Figure 1: *Functional Block Diagram of CTECC Feature Extraction. After [12].*

nent real-valued signal, $x(t)$, TEO is defined as [13]:

$$\Psi[x(t)] = [\dot{x}(t)]^2 - x(t)\ddot{x}(t) = A^2 sin^2(\omega) \approx A^2\omega^2, \quad (1)$$

where $\dot{x}$(t) and $\ddot{x}$(t) represents the first derivative ($\frac{d}{dt}$) and second-order derivative ($\frac{d^2}{dt^2}$) of the signal $x(t)$ *w.r.t* time, respectively. Additionally, for signal $x(t)$ the amplitude and angular frequency are represented by A and ω, respectively. From eq. (1), TEO estimates the energy with high time resolution and hence, localized characteristics of the signal can be tracked. However, TEO is implemented for single channel analysis. Hence, to track the cross-teager energies between two channels, CTEO is developed in [14], and can be denoted as $\Psi_{cr}[\cdot]$. CTEO is a nonlinear quadratic operator, which estimates the relative rate of change of energies between signals. The Cross-Teager Energy (CTE) between the two *real-valued* signals, $x(t)$ and $y(t)$ in continuous-time domain is represented as [15]:

$$\Psi_{cr}[x(t), y(t)] = (\dot{x}(t)\dot{y}(t)) - (x(t)\ddot{y}(t)), \quad (2)$$

$$\Psi_{cr}[y(t), x(t)] = (\dot{y}(t)\dot{x}(t)) - (y(t)\ddot{x}(t)). \quad (3)$$

From eq. (2) and eq. (3), the non-commutative property of CTEO is observed, i.e., $\Psi_{cr}[x(t), y(t)] \neq \Psi_{cr}[y(t), x(t)]$ [14], [16]. Using eq. (2), the *average CTEO* ($\Psi_{cr}^{avg}[\cdot]$) between the continuous-time *real-valued* signals is estimated as [16]:

$$\Psi_{cr}^{avg}[x(t), y(t)] = \frac{1}{2}(\Psi_{cr}[x(t), y(t)] + \Psi_{cr}[y(t), x(t)]). \quad (4)$$

However, the definition of CTEO can be extended to complex-valued signals as given in [17]. Furthermore, for the discrete-time signals $x(n)$ and $y(n)$, average cross-Teager energies are estimated as:

$$\Psi_{cr}^{avg}\{x(n), y(n)\} = x(n)y(n) - 0.5[x(n+1)y(n-1) + x(n-1)y(n+1)]. \quad (5)$$

$$\Psi\{x(n)\} = x^2(n) - x(n+1)x(n-1). \quad (6)$$

From eq. (5), the excellent time resolution of the CTEO can be observed. Subsequently, the later part of the paper deals with the real-valued continuous-time domain representation of speech signal, which can be further extended in discrete-time.

Let us consider the signal $x_i(t)$ in N-sensor microphone array, where $i \in [1, N]$ and $x_i(t)$ is represented as:

$$x_i(t) = s_i(t) + n_i(t), i = 1, 2, ..., N, \quad (7)$$

where $s_i(t)$ and $n_i(t)$ represent the original speech signal and additive noise in i^{th} sensor, respectively. The additive noise component is assumed to be zero-mean and Wide Sense Stationary (WSS) Gaussian random process. Further, the output signal of each sensor $x_i(t)$ is decomposed using a suitable filterbank into L subband signals, and subband filtered signal is represented as:

$$x_{i_j}(t) = x_i(t) * g_j(t), \qquad j = 1, 2, ..., L, \quad (8)$$

where '*' represents the convolution and $x_{i_j}(t)$ represents the subband filtered signal obtained for the i^{th} channel and j^{th} subband filter in the filterbank. Considering two sensor input (p, q) and j^{th} subband filter of the filterbank, the CTE will be expressed as:

$$\Psi_{cr}[x_{p_j}(t), x_{q_j}(t)] = (\dot{x}_{p_j}(t)\dot{x}_{q_j}(t)) - (x_{p_j}(t)\ddot{x}_{p_j}(t)). \quad (9)$$

From the eq. (1), eq. (7), and eq. (9), we obtain:

$$\Psi_{cr}[x_{p_j}(t), x_{q_j}(t)] = \Psi_{cr}[s_j(t)] + \Psi_{cr}[n_{p_j}(t), n_{q_j}(t)] + \Psi_{cr}[s_j(t), n_{q_j}(t)] + \Psi_{cr}[n_{p_j}(t), s_j(t)]. \quad (10)$$

The additive noise represented by the last three terms on the Right-Hand Side (RHS) of eq. (10). Applying taking expectation operator ($E[\cdot]$) on eq. (10), we get:

$$E\{\Psi_{cr}[x_{p_j}(t), x_{q_j}(t)]\} = E\{\Psi_{cr}[s_j(t)]\} + E\{\Psi_{cr}[n_{p_j}(t), n_{q_j}(t)]\}. \quad (11)$$

The last two terms of RHS side in eq.(10) are zero-mean and hence, the expectation operator is zero [15]. However, the second term represents the error in eq. (11) [18]. Hence, the modified equation is given as:

$$E\{\Psi_{cr}[x_{p_j}(t), x_{q_j}(t)]\} = E\{\Psi_{cr}[s_j(t)]\} + error. \quad (12)$$

Let us denote τ the concentration of noise power within the subband filter's passband. Using Cauchy–Schwartz inequality for two random variables *P* and *Q*, we have [19]:

$$|E(PQ)|^2 \leq E(P^2)E(Q^2), \quad (13)$$

where *(PQ)* is the inner product between the random variables *P* and *Q*. Therefore, using eq. (13), the relation between the noise power, we obtain:

$$|\tau_{(pq)_j}| \leq \tau_{p_j}\tau_{q_j}, \quad (14)$$

where τ_{p_j} is the noise power concentration of the j^{th} subband and p^{th} channel. Moreover, τ_{p_j} is proportional to the *error* term in eq. (12), where the *error* term is the varying, whereas the source signal through the bandpass filter remains the same throughout the analysis. Furthermore, in ASR application, the desirable speech signal representation should contain the least amount of noise component and more linguistic

information [13]. Hence, to capture maximum linguistic information in this study, we have considered minimum *error* for the severity-level classification of dysarthric speech. To that effect, in the proposed CTEO, we have ${}^{N}C_2$ possibilities of channel-pairs for each i^{th} subband. Estimating the lowest average CTE for among all channel-pairs is a feasible, yet, computationally expensive approach. Hence, to increase the computational efficiency, we have selected the two channels with the lowest average Teager Energy (TE) and performed the CTEO of the selected channels. Further, from the set of two TE and one CTE we select the signal with the minimum energy for the classification of severity-level classification of dysarthria namely, Minimum Energy Signal (MES). Mathematically, MES can be represented as:

$$MES = min_{(p,q)}(E\{\Psi_{cr}^{avg}[x_{p_j}(t), x_{q_j}(t)]\}, E\{\Psi_{cr}[x_{p_j}(t)]\}, E\{\Psi_{cr}[x_{q_j}(t)]\}). \tag{15}$$

From eq. (15), the MES contains the maximum linguistic information captures through the CTEO. Lastly, the MES is used to further processing for the severity-level classification of dysarthria.

2.1. CTECC Feature Extraction Procedure

Figure 1 shows the functional block diagram of the proposed CTECC for the designing of the severity-level classification system for dysarthria. The microphone array is utilized for the capturing of the dysarthria speech at a sampling rate of $16kHz$. The input dysarthic speech from the *N*-channel microphone array is processed through Gabor filterbank, which has an excellent time-frequency resolution (because the Fourier transform of a Gaussian function is also a Gaussian) [20]. The Gabor filterbank consist of linearly-spaced 40 subband filters and hence, we obtain 40 subband filtered signals for each channel. Next, for each of the subband filter, the TEO profile is estimated. TEO profile for corresponding to the j^{th} subband filter is estimated for n^{th} microphone array channel represented in Fig. 1 as TEO_{nj}, where $j \in [1, 40]$ and $n \in [1, N]$ channels of microphone array. Windowing is performed on the subband filtered signal with window size of $30ms$ and window shift of $15ms$, which provides m frames. Averaging on each frame is performed, which provides the average energy for a frame in consideration. Then logarithm operation is performed, which is followed by Discrete Cosine Transform (DCT) to obtain the cepstral representation. Initial 40 DCT (static) features are concatenated with dynamic Δ and $\Delta\Delta$ coefficients, which results in *120*-dimensional (*120*-D) CTECC feature set [21].

3. Experimental Setup

3.1. Dataset Used

The proposed CTEO-based technique is evaluated using the Universal Access dysarthric Speech (UA-Speech) corpus [22]. In our experiments, dataset configuration mentioned in [5] was used for baseline evaluation. For CTECC feature extraction, microphone array number $M3$, $M5$, and $M6$ for each speaker were used. Apart from these, 465 word utterances out of 765 utterances were used. For training, we used 90% of data, which comprises 837, 837, 833, and 676 utterances. Similarly, for evaluation of the classification system, 10% of the data is utilized, consisting of total 354 utterances.

Table 1: *Class-wise patient details. After [22].*

	Female	Male	Number of Samples
High	F03	M01, M04, M12	751
Medium	F02	M07, M16	930
Low	F04	M05, M11	926
Very Low	F05	M08, M09, M10, M14	930

3.2. Details of Feature Used

As mentioned in [5], the STFT was applied to generate a time-frequency representation with a window size of 2 ms and window overlap of 0.5 ms.

$$X(\omega, \tau) = \sum_{n=-\infty}^{n=+\infty} x[n].w[n, \tau].e^{-j\omega n}, \tag{16}$$

followed by computation of spectrogram (i.e., $|X(\omega, \tau)|^2$ that is fed as input to CNN classifier. Furthermore, the performance of CTECC feature set is analysed using 120-D feature set, where 40 coefficients are static, 40 Δ coefficients, and 40 $\Delta\Delta$ coefficients. CTECC feature set is extracted using the 40 subband Gabor filters, with the center frequency placed on linear scale.

3.3. CNN Classifiers

The Convolutional Neural Network (CNN) is utilized as a classifier in this study based on the experiments given in [7]. CNN performs similarly to other deep neural network (DNN)-based classifiers for the UA-Speech corpus, according to a study published in [7]. The CNN model was trained using the Adam optimizer algorithm [23], three convolutional layers with kernel sizes of 5×5, and one Fully-Connected (FC) layer [24]. Rectified Linear Activation (ReLU) [25] and a max-pool layer are used. A learning rate of 0.001 and cross-entropy loss are used to estimate loss [26].

3.4. Performance Evaluation

The performance of CTECC feature set was compared against the baseline STFT feature set using various performance evaluation metrics, such as $F1-$ Score, Mathew's Correlation Coefficient (MCC), Jaccard's Index, and Hamming Loss.

3.4.1. F1-Score

It is a widely used statistical parameter for evaluating the performance of a model. It is estimated as the harmonic mean of the model's precision and recall [27]. In particular:

$$F1 - score = \frac{2TP}{2TP + FP + FN}, \tag{17}$$

where TP, FP, and FN, represents True Positive, False positive, and False Negative, respectively. It has a value of 0 to 1, with a score closer to 1 signifying better performance.

3.4.2. Matthew's Correlation Coefficient (MCC)

It shows the degree of association between the expected and actual class [28]. It is usually considered a balanced measure when comparing models. MCC is in the range of -1 to 1. It is given as:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TN + FN)(TP_FN)(TN + FP)}}. \tag{18}$$

3.4.3. Jaccard Index

The Jaccard index is a measure of how similar and dissimilar two classes are. It's value is between 0 and 1. It is described in [29]:

$$\text{Jaccard Index} = \frac{TP}{TP + FP + FN}. \quad (19)$$

3.4.4. Hamming Loss

It takes into account incorrectly predicted class labels. All classes and test data are normalized for prediction error (prediction of an inaccurate label) and missing error (prediction of a relevant label). Hamming loss can be calculated using the formula below [30]:

$$\text{Hamming Loss} = \frac{1}{nL} \sum_{i=1}^{n} \sum_{j=1}^{L} I(y_i^j \neq \hat{y}_i^j), \quad (20)$$

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

The % classification accuracy of baseline STFT and CTECC on CNN is shown in Table 2. It can be observed that the CTECC (min) performs better with classification accuracy of 95.76% than the baseline STFT and CTECC (max) on CNN model. Furthermore, the performance analysis shown in the Table 3 using statistical parameters, such as $F1-$Score, MCC, Jaccard Index, and Hamming Loss, also shows that the CTECC (min) shows better linguist information capturing capabilities from the dysarthric speech compared to CTECC (max) and STFT feature set on CNN model. In addition to it, Table 4 shows the confusion matrix of STFT, CTECC (max) and CTECC (min) feature set. It can be observed from the table that the false prediction is reduced by the CTECC (min) in comparison to baseline STFT and CTECC (max) feature set, which all the more supports the fact that CTECC (min) is capable of capturing the linguist information better than STFT and CTECC (max) feature set.

Table 2: *% Classification Accuracy for Baseline STFT and CTECC Feature Set*

Feature Set	CNN
Spectrogram	91.72
CTECC_max	91.24
CTECC_min	**95.76**

Table 3: *Performance Evaluation for Various Feature Set*

Feature Set	F1-Score	MCC	Jaccard Index	Hamming Loss
STFT	0.87	0.83	0.776	0.124
CTECC_max	0.91	0.88	0.84	0.087
CTECC_min	**0.96**	**0.94**	**0.91**	**0.042**

4.1. Analysis of Latency Period

Finally, we also analysed the latency period for CTECC (Min) and CTECC (Max) feature sets as shown in Figure 2. 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 [31]. For latency period analysis, we chose the duration of the utterances varying from 20 ms to 400 ms. 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 CTECC gave significant % classification accuracy in a short duration of *w.r.t* CTECC_max.

Table 4: *Confusion Matrix for STFT, and CTECC Feature Set*

Feature Set	Severity	High	Medium	Low	Very Low
STFT	**High**	63	6	3	3
	Medium	10	79	3	1
	Low	3	4	79	7
	Very Low	1	2	1	89
CTECC (Max)	**High**	62	10	2	1
	Medium	4	85	1	1
	Low	1	3	88	1
	Very Low	1	4	2	86
CTECC (Min)	**High**	**70**	3	2	0
	Medium	3	**90**	0	0
	Low	1	3	**87**	2
	Very Low	0	1	0	**92**

Hence, these results signifies the suitability of CTECC for deployment of practical dysarthric speech classification system.

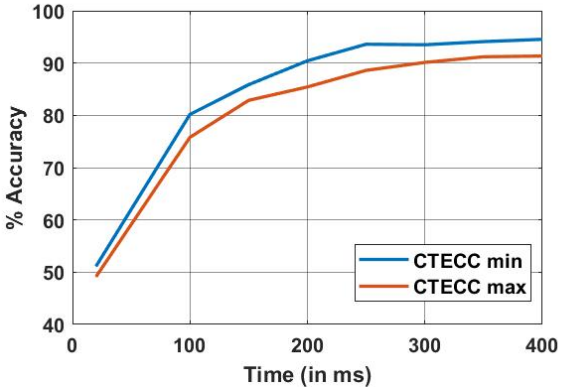

Figure 2: *Latency Period* vs. *% Accuracy Comparison Between CTECC min and CTECC max. After [31].*

5. Summary and Conclusions

In this study, we investigated the discriminative power of CTECC in dysarthric severity-level classification. Further, the effect of microphone arrays are taken into consideration for classifying the discriminative acoustic cues. It can be observed that the CTECC (Min) shows the better classification accuracy of severity-level classification of dysarthria, which signifies the effectiveness of linguistic information that is captured by CTECC_min. Furthermore, we have also demonstrated the discriminative capability of CTECC using statistical measures, such as $F1$-score, MCC, Jaccard index, and Hamming loss. However, the extraction of the CTECC features is computationally expensive. Other dysarthic speech corpora, such as TORGO and Home service, will be used to further validate this work in the future.

6. Acknowledgements

The authors would like to thank the Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, for sponsoring consortium project titled 'Speech Technologies in Indian Languages' under 'National Language Translation Mission (NLTM): BHASHINI', subtitled 'Building Assistive Speech Technologies for the Challenged' (Grant ID: 11(1)2022-HCC (TDIL)). We also thank the consortium leaders Prof. Hema A. Murthy, and Prof. S. Umesh for their support and cooperation to carry out this research work. The authors would like to thank the organizers of UA Speech Corpus for making UA-Speech corpus publicly available. Without these, this work could not have been possible.

7. 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] 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.

[6] B. A. Al-Qatab and M. B. Mustafa, "Classification of dysarthric speech according to the severity of impairment: An analysis of acoustic features," *IEEE Access*, vol. 9, pp. 18 183–18 194, 2021.

[7] 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.

[8] 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.

[9] P. D. Green, J. Carmichael, A. Hatzis, P. Enderby, M. S. Hawley, and M. Parker, "Automatic speech recognition with sparse training data for dysarthric speakers." in *Interspeech*, 2003.

[10] 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.

[11] I. Rodomagoulakis and P. Maragos, "Improved frequency modulation features for multichannel distant speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 841–849, 2019.

[12] R. Acharya, H. Kotta, A. T. Patil, and H. A. Patil, "Cross-Teager energy cepstral coefficients for replay spoof detection on voice assistants," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Ontario, Canada*, 6-11 June 2021, pp. 6364–6368.

[13] 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.

[14] J. F. Kaiser, "Some useful properties of Teager's energy operators," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Minneapolis, MN, USA*, vol. 3, 1993, pp. 149–152.

[15] S. Lefkimmiatis, P. Maragos, and A. Katsamanis, "Multisensor multiband cross-energy tracking for feature extraction and recognition," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, NV, USA*, 2008, pp. 4741–4744.

[16] A.-O. Boudraa, J.-C. Cexus, and K. Abed-Meraim, "Cross ψ b-energy operator-based signal detection," *The Journal of the Acoustical Society of America (JASA)*, vol. 123, no. 6, pp. 4283–4289, 2008.

[17] J.-C. Cexus and A.-O. Boudraa, "Link between cross-Wigner distribution and cross-Teager energy operator," *Electronics Letters*, vol. 40, no. 12, pp. 778–780, 2004.

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

[19] R. Bhatia and C. Davis, "A Cauchy-Schwartz inequality for operators with applications," *Linear Algebra and Its Applications*, vol. 223, pp. 119–129, 1995.

[20] R. Mehrotra, K. R. Namuduri, and N. Ranganathan, "Gabor filter-based edge detection," *Pattern recognition*, vol. 25, no. 12, pp. 1479–1494, 1992.

[21] K. Kumar, C. Kim, and R. M. Stern, "Delta-spectral cepstral coefficients for robust speech recognition," in *2011 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 4784–4787.

[22] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," 2008.

[23] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*. IEEE, 2018, pp. 1–2.

[24] 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.

[25] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[26] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

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

[28] 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.

[29] 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.

[30] 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.

[31] N. J. Shah, M. A. B. Shaik, P. Periyasamy, H. A. Patil, and V. Vij, "Exploiting phase-based features for whisper vs. speech classification," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 21–25.