# Cochlear Filter-Based Cepstral Features For Dysarthric Severity-Level Classification

Siddharth Rathod, Priyanka Gupta, Aastha Kachhi, and Hemant A. Patil
Speech Research Lab, DA-IICT Gandhinagar, Gujarat, India
Email: {siddharth_rathod, priyanka_gupta, aastha_k, hemant_patil}@daiict.ac.in

***Abstract*—Severity-level classification of dysarthria helps in diagnosing a patient and choosing an appropriate course of treatment. This would also aid in redirecting the speech to an appropriate dysarthric Automatic Speech Recognition (ASR), as traditional ASR does not perform well on dysarthric speech. In the recent past, several approaches have been used to study the severity-level classification of dysarthria using state-of-the-art features, such as Short-Time Fourier Transform (STFT) and Mel Frequency Cepstral Coefficients (MFCC). This study investigates novel auditory transform-based Cochlear Filter Cepstral Coefficients (CFCC) features for dysarthric severity-level classification. Three DNN-based classifiers, namely, Convolutional Neural Network (CNN), Light-CNN (LCNN), and Residual Neural Network (ResNet) were employed on UA-Speech Corpus and TORGO corpus. Our proposed CFCC feature set yields an improved classification accuracy of $97.46\%$ ($98.99\%$), $94.92\%$ ($94.97\%$), and $96.66\%$ ($98.93\%$) on UA (Torgo)-corpus using CNN, LCNN and ResNet classifiers respectively. Furthermore, performance metrics, such as the Jaccard index, Matthew's Correlation Coefficient (MCC), $F1$-score, and Hamming loss are used to examine feature discrimination power of CFCC. Finally, latency period of CFCC was also analysed for practical deployment of system.**

***Index Terms*—Dysarthria, UA-Speech Corpus, TORGO Corpus, CFCC, LFCC, MFCC, CNN.**

## I. Introduction

The speech production mechanism in humans requires proper co-ordination between the brain and the muscles that produce intelligible speech [1]. This co-ordination is affected due to disorders, such as cerebral palsy, muscular dystrophy, stroke, brain infection, brain injury, facial paralysis, tongue or throat muscular weakness, and nervous system disorders. In such unfortunate cases, one may suffer from speech disorders, such as dysarthria, stuttering, apraxia, and dysprosody. Out of these, one of the prevalent speech disorders is dysarthria, wherein the dynamic movements of the articulators, and the upper respiratory system are affected. This leads to difficulty in production of natural speech.

The course of treatment for dysarthric patients is recommended based on their severity-level. Thus, there is active research to develop techniques for classifying dysarthric severity [2]. Therefore, the investigation of severity-level of dysarthria aids in determining the course of treatment to be chosen. In the literature, a considerable use of the Short-Time Fourier Transform (STFT) [3], and numerous acoustical parameters to classify the severity-levels of dysarthria has been made [4]. The state-of-the-art feature sets, such as MFCC were used because of their capacity to capture global spectral envelope information [5]. Along with perceptually justified state-of-the-art feature sets, such as the MFCC, glottal excitation source parameters from quasi-periodic sampling of the vocal tract system were implemented in [6]. Speech signals are non-stationary signals, particularly due to the dynamic range of the multi-frequency components present in them. During natural speech production, the dynamic motion of articulators causes the frequency content to alter rapidly. This rapid change of frequency content is all the more vivid in dysarthric patients, due to irregular movement of the articulators, which also depends on the severity-level of dysarthria.

Furthermore, the desired performance is not obtained, when the acoustic training and testing environments are mismatched. Due to the observed resilience of the human hearing system to mismatched conditions [7], we propose a feature extraction technique based on the fundamental signal processing processes in the ear, utilizing Auditory Transform (AT). An auditory-based time-frequency transform, is the foundation for Cochlear Filter Cepstral Coefficients (CFCC) [8]–[10]. In this work, we propose CFCC to classify the severity-level of dysarthria, made the following contributions in this paper:

- We propose the use of auditory transform-based features, namely, CFCC, for the severity-level classification of dysarthric speech.
- We have experimentally analysed the effects of the cochlear filter parameters, namely, $\alpha$ and $\beta$, on the performance of the classifier used.
- Analysis of impact of frequency resolution on the performance of the classifiers has been done through varying the number of subband filters.
- Since the works on dysarthria requires high performance, it is important to analyze the precision for retraining of the model. The experiments for the same are reported in this work.
- Given that the course of treatment for dysarthria is determined based on the accuracy of the diagnosis of the severity-level, it is important to be able to accurately diagnose the disease even for shorter durations of speech. Hence, we have also included the latency period analysis and also compared it with the state-of-the-art feature sets.

## II. Cochlear Filter Cepstral Coefficients (CFCC)

The CFCC feature set extraction steps for dysarthric severity-level classification are described in this Section.

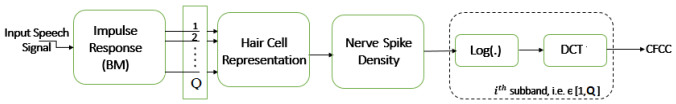


Fig. 1. Functional Block Diagram of the Proposed CFCC Feature Set. After [**?**], [9].

*1) Cochlear Filtering:* There are three main regions of human ear, namely, the outer ear, the middle ear, and the inner ear. The outer region of the human ear consists of the visible part of the ear called *pinna*, which captures sound waves that are funneled into the middle ear. The sound wave is converted to mechanical energy by three bones located in the middle ear. Out of the three bones, the last bone is called as *stapes*. Stapes sets the fluid in cochlea into motion resulting into waves in the Basilar Membrane (BM). In order to decompose the input speech signal to the cochlea into a set of subbands signals, the Auditory Transform (AT) is used. The AT $W(a,b)$ of a signal $s(t)$ w.r.t. a cochlear filter $\psi(t)$ is given by [9]:

$$\begin{aligned} W(a,b) &= \langle s(t), \psi_{a,b}(t) \rangle, \\ &= \int_{-\infty}^{+\infty} s(t) {\psi^*}_{a,b}(t) dt, \\ &= \int_{-\infty}^{+\infty} s(t) \frac{1}{\sqrt{a}} \psi^* \left( \frac{t-b}{a} \right) dt, \end{aligned} \tag{1}$$

where $*$ and $< . >$ denote complex conjugate and inner product operation, respectively. The wavelet transform involves two parameters, $a$ and $b$, both of which are real-valued, where $a \in R^+$ and $b \in R$. The $a$ parameter determines the scale and dilation of the mother wavelet. The center frequency of the impulse response function depends on the value of $a$ being chosen. Multiple values of $a$ are chosen to generate subband filters having desired center frequencies, forming up the filterbank. The parameter $b$ determines the time shift of the wavelet and shift the wavelet by amount $b$ along the time-axis. As an example, Fig. 2 (a) shows the time-domain representation of one of the baby wavelets with a specified center frequency, and Fig. 2 (b) shows frequency response of 10 such baby wavelets each with a different center frequency. The number of subbands is determined by the number of subband filters in a filterbank.

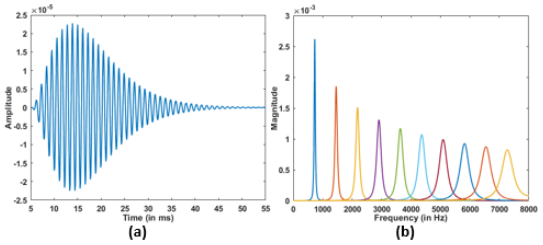


Fig. 2. (a) Impulse Response of $2^{nd}$ Subband (Cochlear) Filter, and (b) Frequency Response of Cochlear Filterbank Comprising Ten Subband Filters.

*2) Hair Cell Representation:* Due to the waves generated by *stapes* in the BM, the hair cells in the cochlea are displaced in *one* direction. This unidirectional hair movement results in neural excitation. The neural excitation ceases if the hair cell displacement is in the opposite direction. This motion of hair cells can be expressed mathematically as [10]:

$$H(a,b) = W(a,b)^2; \quad \forall\, W(a,b), \tag{2}$$

where $W(a,b)$ is the filterbank output.

*3) Nerve Spike Density (NSD):* In addition, the hair cell output for each subband is quantified as an NSD count. This count is obtained by estimating the average of the hair cell output within each frame, which is of length 20 ms and with a shift of 8 ms. For the $i^{th}$ subband and $j^{th}$ frame, the NSD count is computed as follows:

$$NSD(i,j) = \frac{1}{l} \sum_{b=n}^{n+l-1} H(i,b), \quad n = 1, N, 2N, ...; \; \forall i, j, \tag{3}$$

where the $l$ denotes the frame length, $b$ is the sample number, and the frame duration is denoted by $N$. Overall, the feature extraction process consists of selecting $Q$ center frequencies from the frequency range of $[0, \frac{Fs}{2}]$, divided into Q frequencies evenly-spaced, which would be the center frequencies of our subband filters. A series of values of the parameter $a$ are chosen accordingly to generate the wavelet function $\psi(t)$, keeping the value of $b$ = 0. $W(a,b)$ is generated for the given signal $s(t)$ for each subband. Then, the hair cell and nerve spike density of the subband outputs are calculated. Finally, in order to decorrelate the feature dimension and compaction of energy, Discrete Cosine Transform (DCT) is applied to extract the CFCC feature vector. Fig. 1 shows the functional block diagram representation of the proposed CFCC feature set, where 14 static coefficients are taken as input features along with $\Delta$ and $\Delta\Delta$ features to form 42-D feature vector.

Fig. 3 shows the comparative analysis of the four levels of dysarthric severity for the word 'yes' taken from the UA corpus. To that effect, spectrographic analysis is shown in Panel-I, followed by the contour plots of the proposed CFCC feature set in Panel-II. It can be observed that as opposed to the spectrogram, the CFCC feature set shows more discriminating information, as the severity-level increases. In particular, the density of the isolines of the contour plots shown in Panel-II can be observed to be the highest for high severity-level class.

## III. Experimental Setup

### A. Datasets Used

In this work, we use the Universal Access Dysarthric Speech (UA-Speech) corpus [11]. Each speaker's microphone arrays $M3$, $M5$, and $M6$ were used for the extraction of CFCC features. Apart from this, 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 was used, consisting of total 354 utterances. We have also made use of the TORGO dataset taking a total of 1982 utterances from all classes out of which 10% is used for testing [12].

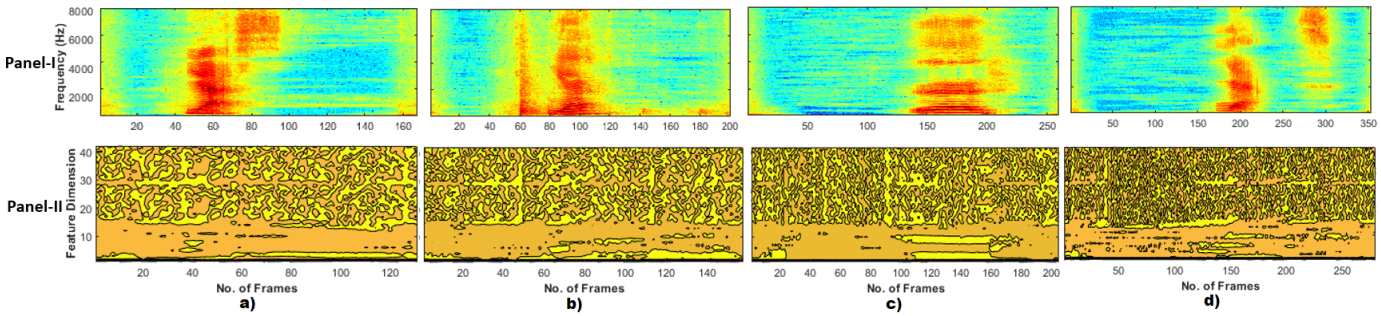


Fig. 3. Comparative analysis of dysarthric severity-levels captured by Panel I: spectrogram *vs.* Panel II: CFCC feature set. For severity-level of (a) very low, (b)low, (c) medium, (d) high.

TABLE I
CLASS-WISE PATIENT DETAILS

| | UA Corpus [11] | TORGO [13] |
|---|---|---|
| High | F03, M01, M04, M12 | - |
| Medium | F02, M07, M16 | M01, M04 |
| Low | F04, M05, M11 | F01, M05 |
| Very Low | F05, M08, M09, M10, M14 | F04, M03 |

### B. Feature Sets Used

In this study, the performance of CFCC is compared with STFT as baseline as in [3]. Further, the performance is also compared with the state-of-the-art feature sets, such as MFCC and Linear Frequency Cepstral Coefficients (LFCC). The parametric details of these features are given in Table II.

TABLE II
DETAILS OF PARAMETERS OF THE VARIOUS FEATURE SETS USED

| **Parameters** | **STFT** | **MFCC** | **LFCC** |
|---|---|---|---|
| Frequency Scale | Linear | Mel | Linear |
| Subband Filter | - | 40 | 40 |
| Window Length | 2 ms | 25ms | 15ms |
| Window Shift | 0.5 ms | 25ms | 15ms |
| Feature Dimension | $512 \times 512$ | 42 | 120 |

### C. Classifiers Used

In this work, initial experiments are reported on Convolutional Neural Network (CNN) classifier [5]. The CNN model used in this work consists of four convolutional layers, and one fully-connected (FC) layer. The kernel size is set to be $3 \times 3$, $4 \times 4$, $5 \times 5$, and $5 \times 5$ for each layers, respectively. Rectified Linear Activation (ReLU) [14] and a max-pool layer are used. The CNN model was trained using the Stochastic-Gradient-Descent (SGD) optimizer algorithm [15]. A learning rate of $0.003$ and cross-entropy loss are used to estimate loss [16]. In order to reinforce our findings, we conducted similar experiments utilizing ResNet and LCNN classifier architectures with comparable depth, while maintaining a consistent learning rate for SGD.

### D. Performance Evaluation

In this work, various performance evaluation metrics, such as the $F1-$ Score, which is a widely used statistical method for evaluating a model's performance [17], Mathew's Correlation Coefficient (MCC), which shows the degree of association between the expected and actual class [18], Jaccard's Index, which measures the similarity and dissimilarity of two classes [19], and Hamming loss, which takes into account the classes that are inaccurately predicted are used [20].

## IV. EXPERIMENTAL RESULTS

### A. Fine-Tuning Cochlear Filter Parameters

This Section shows the various experimental results on UA corpus, using CNN as the classifier, w.r.t. fine-tuning of the cochlear filter parameters, namely, $\alpha$, $\beta$, and the number of subband filters $Q$. The parameter fine-tuning involved varying of one parameter while keeping the other two parameters constant. To that effect, first the number of subband filters (Q) was varied from $40$ to $160$ in the steps of $20$, while keeping $\alpha$ = 3 and $\beta$ = 0.02. Fig. 4 shows the results obtained by varying the number of subband filters, and therby implicitly varying the frequency resolution. It can be observed that the highest performance is achieved for $120$ number of subbands.

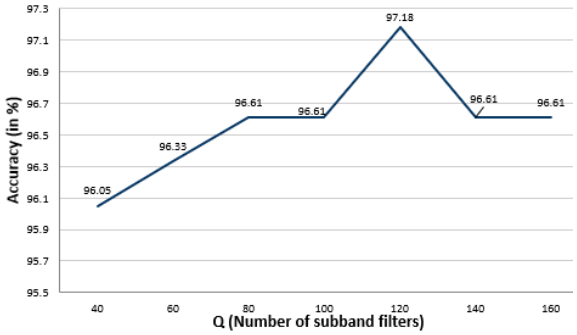


Fig. 4. Effect of Number of Subband Filters in Cochlear Filterbank.

The next set of experiments is done by varying the cochlear filter parameter $\alpha$, and keeping $\beta$ and $Q$ as $0.02$ and $120$, respectively. To that effect, Figure 5 shows the performance when $\alpha$ is varied from $2$ to $4$ in the steps of $0.5$. It can be observed that for $\alpha = 3$, we obtain the highest performance, which has also been the optimal value of $\alpha$ for various other applications of cochlear filter based feature sets [21], [22].

Furthermore, the parameter $\beta$ was varied from $0.016$ to $0.024$ in the steps of $0.002$, keeping $Q$ = 120 and $\alpha$ = 3.

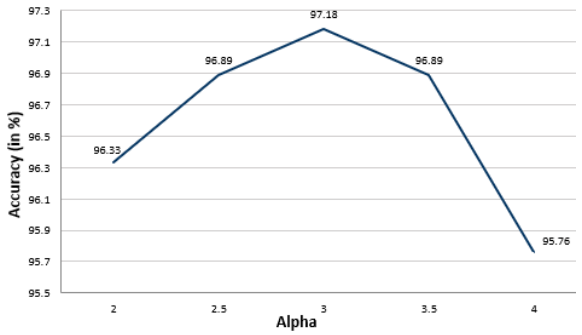


Fig. 5. Effect of Parameter $\alpha$

The corresponding results are shown Fig.6, where the optimal value of $\beta$ obtained is $0.018$, giving the accuracy of $97.46\%$.

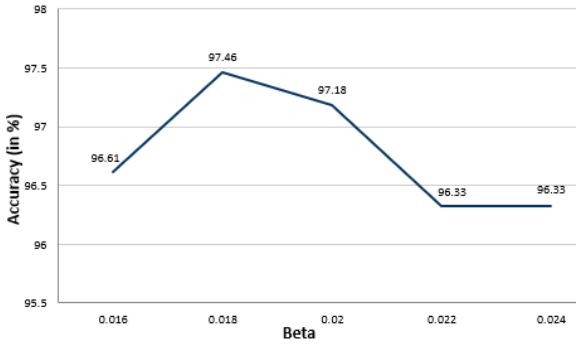


Fig. 6. Effect of Parameter $\beta$

## B. Comparison with Existing Feature Sets

For further experiments, we have used the CFCC features with these optimal parameter values and compared it with the existing features, such as STFT, MFCC, and LFCC in Table III and Table IV. It can be observed that CFCC has lesser false predictions compared to the other *non-auditory transform*-based features. Given that the CFCC feature set is specially designed to mimic the human auditory system by exploiting auditory transform-based filtering, it is more susceptible to the changes and variations caused in speech [7]. Furthermore, the formant structure of dysarthric speech is often distorted due to neuromotor coordination issues, making acoustic signals better captured by cochlear modelling. Therefore, the CFCC feature set is able to capture the unique characteristics of dysarthric speech necessary for dysarthric severity-level classification (such as pitch ($F_o$), loudness, and articulation) and thus, performs better than other non-auditory transform-based feature sets.

TABLE III
PERFORMANCE EVALUATION FOR VARIOUS FEATURE SETS

| Feature Set | Accuracy | F1-Score | MCC | Jaccard Index | Hamming Loss |
|---|---|---|---|---|---|
| **STFT** | 91.53 | 0.87 | 0.83 | 0.78 | 0.124 |
| **MFCC** | 95.20 | 0.91 | 0.88 | 0.84 | 0.087 |
| **LFCC** | 96.05 | 0.96 | 0.96 | 0.93 | 0.034 |
| **CFCC** | **97.46** | **0.97** | **0.97** | **0.95** | **0.025** |

TABLE IV
CONFUSION MATRIX OF BASELINE MFCC, LFCC, AND CFCC

| **MFCC** | **High** | **Medium** | **Low** | **Very Low** |
|---|---|---|---|---|
| **High** | 70 | 2 | 2 | 1 |
| **Medium** | 1 | 88 | 3 | 1 |
| **Low** | 1 | 1 | 88 | 3 |
| **Very Low** | 1 | 1 | 0 | 91 |

| **LFCC** | High | Medium | Low | Very Low |
|---|---|---|---|---|
| High | 68 | 4 | 3 | 0 |
| Medium | 2 | 88 | 2 | 1 |
| Low | 0 | 2 | 91 | 0 |
| Very Low | 0 | 0 | 0 | 93 |

| **CFCC** | High | Medium | Low | Very Low |
|---|---|---|---|---|
| High | 72 | 1 | 1 | 1 |
| Medium | 2 | 88 | 3 | 0 |
| Low | 0 | 1 | 92 | 0 |
| Very Low | 0 | 0 | 0 | 93 |

## C. Effect of Classifier

To further evaluate the robustness of the CFCC feature set we have tested it on three architectures, namely CNN, LCNN, and ResNet.. It is observed in Fig.7 (a) that on the UA corpus, the CNN architecture performs the best for *all* the feature sets. Furthermore, the proposed CFCC feature set outperforms the existing features on all the three architectures. For the case of the TORGO corpus as shown in Fig.7 (b), similar behaviour is observed except for the LCNN architecture.

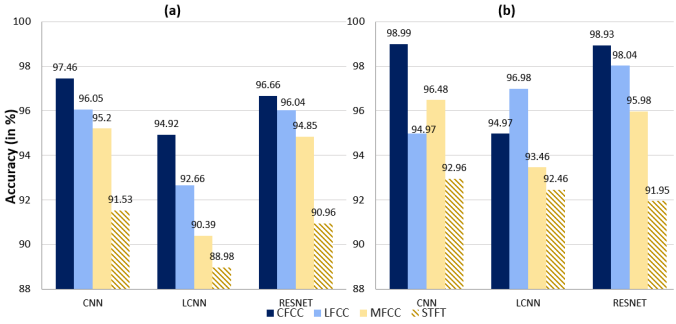


Fig. 7. Effect of classifier using two datasets, namely a) UA Speech Corpus, and b) TORGO Corpus.

## D. Analysis of Precision

Furthermore, to analyze the precision for retraining of the model, experiments were conducted using the CNN model on the datasets, the results remained consistent across 5 trials. A

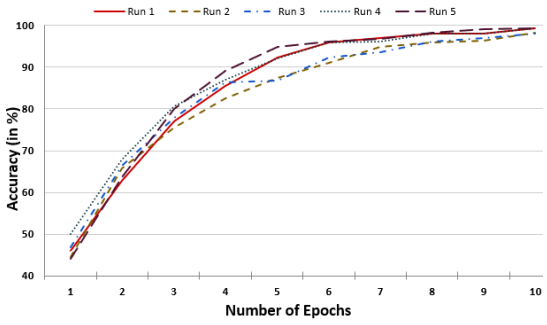


Fig. 8. Analysis of precision for retraining. Best viewed in colour

minor variation in accuracy was observed, which may be due to randomness in the DNN and seed values, as shown in Fig.8. This indicates that the models are robust to quoted precision, should they be *retrained*.

### *E. Analysis of Latency*

In this study, performance of CFCC was also investigated against STFT, MFCC, and LFCC by analysing the latency period as showed in Figure 9. Latency period was estimated by evaluating % classification for varying test speech segment duration. Test segment of 500 to 3000 ms was considered for analysis of latency period. It is evident from figure 9 that CFCC features gave increased % classification accuracy, for speech segment as small as 800 ms, however other features sets gave increased % classification accuracy for speech segments of duration $> 1500$ ms. In the context of these results, the sustainability of CFCC for deployment of practical system is evident.

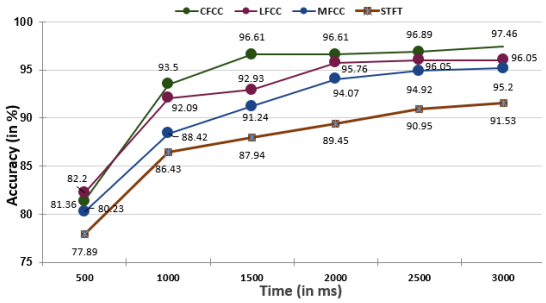


Fig. 9. Analysis of Latency Period for Various Feature Sets.

## V. Summary and Conclusions

This study represents the first detailed analysis of CFCC for dysarthric severity-level classification. To that effect, auditory transform-based CFCC feature set is proposed for classifying the severity-level of dysarthric speech on two corpora, UA-Speech and TORGO. The performance is evaluated using three different classifiers: CNN, LCNN, and ResNet. The CFCC feature set is observed to outperform the existing non-auditory transform-based features, such as MFCC and LFCC. In addition to this, analysis of precision is done, to evaluate the robustness of the system towards the quoted precision, if the model is retrained. Furthermore, we also analysed the latency period *w.r.t.* MFCC and LFCC, which indicates the potential of CFCC for practical deployment of severity-level classification system. In future, the performance of the proposed feature set can be evaluated under cross-database scenarios.

## VI. Acknowledgment

Authors present sincere gratitude to MeitY, New Delhi, Govt. of India, for the project 'Speech Technologies in Indian Languages BHASHINI', (Grant ID: 11(1)2022-HCC (TDIL)) for their support.

## 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] P. Balzan, C. Tattersall, and R. Palmer, "Non-invasive brain stimulation for treating neurogenic dysarthria: A systematic review," *Annals of Physical and Rehabilitation Medicine*, vol. 65, no. 5, p. 101580, 2022.

[3] S. Gupta *et al.*, "Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments," *Neural Networks*, vol. 139, pp. 105–117, 2021.

[4] 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.

[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. 3127–31.

[7] Shao, Yang and Jin, Zhaozhang and Wang, DeLiang and Srinivasan, Soundararajan, "An auditory-based feature for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, Taiwan*, 2009, pp. 4625–4628.

[8] Q. Li, "Solution for pervasive speaker recognition," *Submitted to NSF IT.F4, SBIR Phase I Proposal*, 2003.

[9] Li, Qi, "An auditory-based transfrom for audio signal processing," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York*, 18-21 October, 2009, pp. 181–184.

[10] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Transactions on Audio, Speech, and Language processing*, vol. 19, no. 6, pp. 1791–1801, 2010.

[11] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," *INTERSPEECH, Brisbane, Australia*, pp. 1741–1744, 2008.

[12] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, USA*, 2017, pp. 5070–5074.

[13] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, pp. 523–541, 2012.

[14] A. F. Agarap, "Deep learning using rectified linear units (relu)," *CoRR*, vol. abs/1803.08375, 2018, {Last Accessed: Feb 6, 2023}. [Online]. Available: http://arxiv.org/abs/1803.08375

[15] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th IWQoS*. IEEE, 2018, pp. 1–2.

[16] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in NIPS, Montréal, Canada*, vol. 31, 2018.

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

[18] 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.

[19] 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.

[20] 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.

[21] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral, and instantaneous frequency features for detection of natural *vs.* spoofed speech," in *INTERSPEECH, Dresden, Germany*, 6-10, September 2015, pp. 2062–2066.

[22] P. Gupta, P. K. Chodingala, and H. A. Patil, "Replay spoof detection using energy separation based instantaneous frequency estimation from quadrature and in-phase components," *Computer Speech & Language*, vol. 77, p. 101423, 2023.