Dysarthria Severity Classification Using Phase Based Features of LP Residual

Rohini Sri Mannepalli[1] , Aditya Pusuluri[2] and Hemant A.Patil[3]
Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana
E-mail: 2210040095@klh.edu.in
Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat
E-mail:{aditya_pss,hemat_patil}@daiict.ac.in

***Abstract*—Classifying the severity of speech impairment due to dysarthria is crucial for optimizing care and enhancing communication abilities for affected individuals. This study explores the use of the Modified Group Delay Function (MGDF) of LP residual signal in classifying dysarthria severity-levels. Evaluations were conducted using standard UA-Speech and TORGO datasets. A stratified Convolutional Neural Network (CNN) with 5-fold cross-validation validated the results. Baseline features included Linear Frequency Cepstral Coefficients (LFCC), Mel Frequency Cepstral Coefficients (MFCC), and Whisper module. Key performance evaluation metrics were accuracy, precision, recall, and F1-score. Finally, the latency period was analyzed for practical deployment of the system, system's ability to accurately recognize and process speech from any speaker, without needing to be specifically trained or adapted to individual voice characteristics.**

Keywords: Dysarthria, LP residual, Modified Group Delay Function.

I. Introduction

Dysarthria is a motor speech disorder characterized by impaired movement of the muscles used for speech, resulting in slurred, slow, or difficult-to-understand speech. This condition presents significant challenges in clinical assessment and management due to its heterogeneous nature and varying dysarthric severity-levels. Accurate classification of dysarthria severity is essential for guiding treatment planning and monitoring disease progression. Traditional methods of severity classification, often relying on subjective analysis, are expensive and time-consuming, highlighting the need for more efficient and objective approaches [1].

Early approaches in dysarthria classification utilized acoustic features derived from fundamental or pitch frequency (F_o), formant frequencies, and duration measures [2]. However, these traditional features often lack the sensitivity and specificity required for precise severity-level classification, particularly in subtle or nuanced speech impairments. Researchers have found cepstral-based features, such as Mel Frequency Cepstral Coefficients (MFCC), [3] effective for classifying dysarthria severity due to their ability to capture the vocal tract system characteristics. Studies have shown that combining MFCC with auditory features yields better classification results.

This study proposes the use of Modified Group Delay Cepstral Coefficients (MGDCC) for dysarthric severity-level classification [4], leveraging the phase information of linear prediction (LP) residual to capture the characteristics of dysarthric speech. By integrating features, such as MFCC, Linear Frequency Cepstral Coefficients (LFCC), and Linear Frequency Residual Cepstral Coefficients (LFRCC) with MGDCC, [5] the proposed approach aims to enhance the reliability and accuracy of automated dysarthria severity-level classification systems.

The organization of the rest of the paper is as follows as: Section II presents details of proposed phase-based features from LP residual. Section III gives details of experimental setup used for this study. Section IV presents experimental results evaluating the proposed features with various performance evaluation factors. Finally, Section V concludes the paper along with potential future research directions.

II. Proposed Methodology

A. LP Residual

Speech signal is a one-dimensional signal carrying information in the form of dependencies in the sequence of samplesof speech signal and thus, a single speech sample has no perceputal significance. Hence, LP analysis models speech signals as the result of a linear combination of past samples, aiming to predict the current speech sample. The residual represents the difference between the actual signal and the predicted signal. This residual signal contains information about the components of the speech signal that are not well-predicted by the LP model, typically capturing aspects, such as aspiration noise in consonants, glottal pulses in voiced speech, and other high frequency components [6] . In particular,

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k). \tag{1}$$

The parameter p, which denotes the filter order, significantly influences performance in speech recognition tasks, with optimal results achieved using LP order 10. Residual error plays a crucial role in validating and debugging LP models, guiding iterative optimization algorithms, and refining solutions to improve their quality. By analyzing residual errors, one can identify and correct issues within the model, ensuring more accurate and reliable outcomes. It is given by:

$$r(n) = s(n) - \hat{s}(n) \tag{2}$$

LP residual is a versatile tool in speech signal processing, valued for its ability to capture prediction errors and excitation source details that contribute to the accurate representation and analysis of speech signals in various applications.

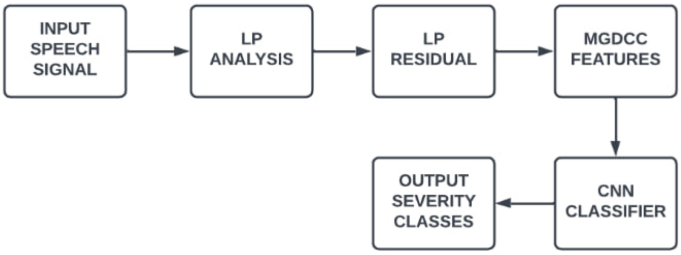

Fig. 1: Functional block diagram of LP residual and MGDCC feature for the dysarthric severity-level classification system.

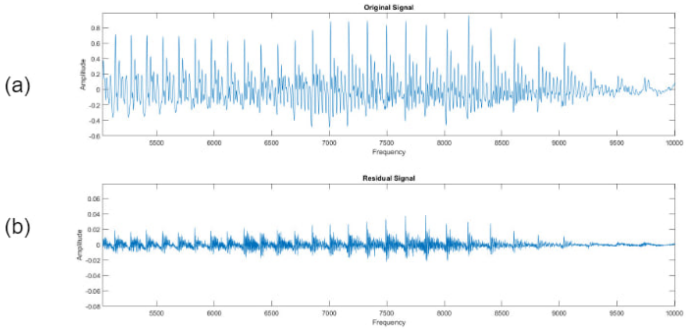

Fig. 2: (a) depicts original signal and (b) depicts MGDCC(lp) of that signal

B. Modified Group Delay Function

Speech signals are complex, characterized by fundamental components, such as frequency, amplitude, and spectral envelope. The vocal tract system, which is generally a maximum phase system, plays a crucial role in shaping these signals by influencing acoustic resonances, known as *formants*, which are essential for distinguishing vowels and certain consonants [7]. The glottal excitation source, with its nonlinear vibration, contrasts with the vocal tract's minimum phase behavior. Analyzing phase provides insights into the temporal coherence and stability of speech components, aiding in tasks, such as speech synthesis, enhancement, and recognition [8]. Identifying resonance frequencies or formant frequencies from the phase spectrum presents a challenge because they become obscured by the phase wrapping phenomenon occurring at multiples of 2π. To overcome this signal processing conflict, the signal must be a minimum phase signal, where the continuous (i.e., unwrapped) phase function is denoted by $\theta(e^{j\omega})$. Minimum phase signals are preferred because their magnitude spectrum and group delay spectrum exhibit similar α characteristics. The group delay function derived as the negative derivative of the unwrapped Fourier transform phase, serves as a measure to quantify this *coherence*. The group delay function is given as,[8]

$$T(e^{j\omega}) = -\frac{d\theta(e^{j\omega})}{d\omega}. \tag{3}$$

The group delay function is specifically applicable to minimum phase signals. However, speech signals are generally mixed-phase systems that contain zeros introduced by noise or nasal sounds. To address these undesirable or spurious spikes, the Modified Group Delay Function (MODGF) is introduced, which effectively mitigates the influence of these zeros near the unit circle in the group delay spectrum of speech signals. The Modified Group Delay Function (MODGF) effectively moves the zeros radially inside the unit circle, thereby reducing the occurrence of spikes in the valleys. Additionally, a cepstrally-smoothed signal is introduced to restore the dynamic range and mitigate the spiky structure of the phase-based features.The modified group delay function is given as:

$$T_m(\omega) = \frac{T(\omega)}{|T(\omega)|}[T(\omega)]^a, \tag{4}$$

where $T(\omega)$ is given by,

$$T(\omega) = \frac{G_R(\omega)H_R(\omega) + G_I(\omega)H_I(\omega)}{|S(\omega)|^{2\gamma}}, \tag{5}$$

where $S(\omega)$ represents the cepstrally-smoothed version of $G(\omega)$, and G_R, G_I, H_R, and H_I indicate the real and imaginary parts, respectively. Two parameters α and γ are introduced , which are used to restore the dynamic range and reduce the amplitude of the unwanted spikes, respectively. The range of α and γ range between $0 < \alpha \leq 1$ and $0 < \gamma \leq 1$.

III. Experimental Setup

A. Dataset Used

This study utilizes two well-known dysarthria speech corpora, namely, the Universal Access Dysarthria Speech Corpus (UA-Speech) and TORGO. Both corpora predominantly exhibit spastic dysarthria, characterized by features, such as breathiness, hypernasality, a harsh voice, and incorrect articulation, which lead to unintelligible speech [9]. The UA-Speech corpus is segmented into four severity-levels: 930 samples of very low severity, 926 samples of low severity, 930 samples of medium severity, and 751 samples of high severity. It includes 8 speakers: 4 males (M01, M05, M07, and M09) and 4 females (F02, F03, F04, and F05). The TORGO corpus comprises a total of 1982 samples distributed across three severity-levels: 671 samples of very low severity, 627 samples of low severity, and 684 samples of medium severity [10]. In both corpora, 80% of the data is allocated for training purposes, while the remaining 20% is reserved for testing. The division ensures that both training and testing sets include words, non-words, and sentences. The experiments employ a *5* cross-validation (CV) approach, focusing on evaluating model performance exclusively on the training data to assess its robustness and speaker-independence.

B. Classifier Used

The research employed a Convolutional Neural Network (CNN) classifier due to its inherent translation invariance, essential for recognizing features irrespective of their *spatial* location. Moreover, CNNs efficiently capture spectro-temporal patterns and variations in speech signals, making them ideal for robust and accurate classification in such audio processing tasks. The model was trained using a stratified *5* cross-validation approach with a set seed value to ensure consistent data distribution across folds. Each fold involved an 80% training set and a 20% validation set. The adam optimizer was employed with categorical cross-entropy as the loss function and accuracy as the evaluation metric. A grid search was performed to optimize the learning rate and batch size over 100 epochs, ensuring that the model's parameters were fine-tuned for peak performance. Two activation functions were utilized: ReLU and softmax. ReLU enhances learning speed and reduces computational cost, applied throughout except at the final layer, where softmax aids multi-class classification. Each convolutional layer incorporated normalization and dropout layers to curb overfitting. Fine-tuning parameters yielded a learning rate of 0.001, batch size of 128, and 100 epochs.

C. Baseline Features

The baseline for this work includes features, such as MFCC, LFCC, LFRCC, and WSPSR (Web- scale Supervised Pretraining). Additionally, the Whisper Tiny model **radford2023robust**, is used as a baseline for the task of dysarthria severity-level classification. This Tiny model is the smallest variant, featuring relatively a fewer trainable parameters and layers compared to its counterparts. For both datasets, LFCC features outperform MFCC features, highlighting the effectiveness of the linear frequency scale for classifying dysarthria severity.

IV. Experimental Results

This Section evaluate the proposed LP residual MGDCC feature set for various experimental evaluation factors, such as the parameter tuning, spectrographic analysis, comparision of baseline features and analysis of latency period.

A. Spectrographic Analysis

The spectrograms collectively illustrate varying degrees of spectral complexity via pattern of spectral energy distribution. Exhibit predominantly low-frequency content with a minimal variation, while the others display a richer spectrum with multiple frequency components. In particular, reveals a wide band of frequencies, indicative of a potentially noisy or intricate signal.

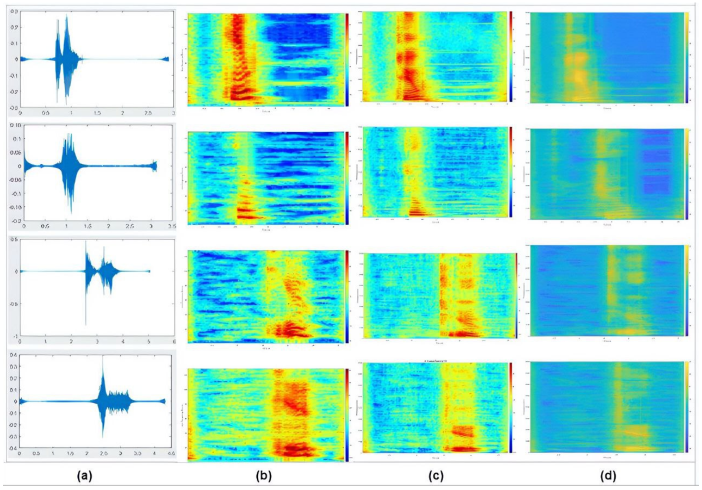

Fig. 3: Plots for dysarthric speech ”to” from UA-Speech:Fig. 3(a), Fig. 3(b), Fig. 3(c), Fig. 3(d), of each section depicts the time-domain waveform, Mel Spectrogram, residual spectrogram, and modified group dealy gram, respectively.

B. Parameter Tuning

As discussed in Eq. (4) and Eq. (5), MGDCC consists of two parameters, namely, α and γ that are fine-tuned using a greedy search algorithm and are varied within the range [0.1,1] with a step size of 0.1. The evaluation is performed using a CNN classifier, and a *5*-fold cross-validation accuracy metric. MGDCC features that are tuned for α = 0.1 and γ = 0.1 results in relatively optimum performance for both UA-Speech and TORGO with fold (test) accuracies of 94.63% (93.45%), and 94.63% (93.37%), respectively, indicating α and γ are the generalized parameters for the dysarthria severity-level classification task w.r.t Eq. (4) and Eq. (5).

C. Comparison with Baseline Features

Table 1 shows the relative comparision of LFRCC feature set along with baseline features, such as MFCC, LFCC, and state-of-the-art whisper model-based features using CNN classifier.

For both the datasets, LFCC outperforms MFCC features, indicate effectiveness of linear frequency scale for the task of dysarthria severity-level classification. LFRCC outperforms both MFCC and LFCC baseline features by a fold (test) accuracy margin of 4.72% (4.23%), 0.51% (1.98%) for UA-Speech, and 5.72%(2.81%), 5.58% (2.38%) for TORGO. Furthermore, state-of-the-art web-scale supervised pre-training for speech recognition (WSPSR), also known as Whisper encoder module is used as a baseline for dysarthria severity-level classification task. The proposed LFRCC features

TABLE I: Fold (test) Accuracy, Precision (P), Recall (R) for Various Feature Sets using CNN classifier on UA-Speech and TORGO

Data	Features	Fold Acc.	Test Acc.	Precision	Recall
UA-Speech	MFCC	87.29	90.68	91.52	90.21
	LFCC	91.50	92.93	93.61	92.41
	Whisper	92.01	94.80	93.28	92.44
	LFRCC	92.01	**94.91**	**95.06**	94.85
	Proposed Features MGDCC (LP)	**93.07**	94.63	94.35	**94.65**
TORGO	MFCC	85.71	88.62	89.40	88.64
	LFCC	90.58	91.82	91.51	91.02
	Whisper	90.97	94.10	93.47	94.00
	LFRCC	91.43	**94.20**	93.96	**94.19**
	Proposed Features MGDCC (LP)	**94.95**	93.45	**94.16**	93.51

perform on par for UA-Speech and outperforms by fold (test) accuracy of 0.46% (0.2%) for TORGO, when compared with whisper model, which is an advance machine learning model trained using labelled data of 680,000 hours. Table I indicate the classwise precision, recall, and F1-Score of LFRCC with p = 10 for UA-Speech and TORGO database. We notice the balanced Fl-score across all the severity-levels reflects the model's robustness and consistency in classification performance.

D. Analysis of Latency Period

Analyzing the latency period helps us to identify the minimum number of speech frames, i.e, minimal speech duration required to achieve optimal classification accuracy. By understanding the relationship between the number of frames and the accuracy, we can fine-tune our models to be both efficient and effective, ensuring that we do not process more data than needed while still maintaining high accuracy in our speech classification tasks. Fig.3 the performance metrics (fold accuracy, accuracy, precision,recall, and F1 score) against different latency periods (50,100, 200, 300, 400, 500, 1000, 1500, 2000ms).Observing the output plot allows one to see how these performance metrics change as latency increases. Observing the output plot allows one to see how these performance metrics change as latency increases, reflects how responsive the proposed features are. Typically, it can highlight the trade-offs between quicker decision-making (lower latency) and the performance of the model.

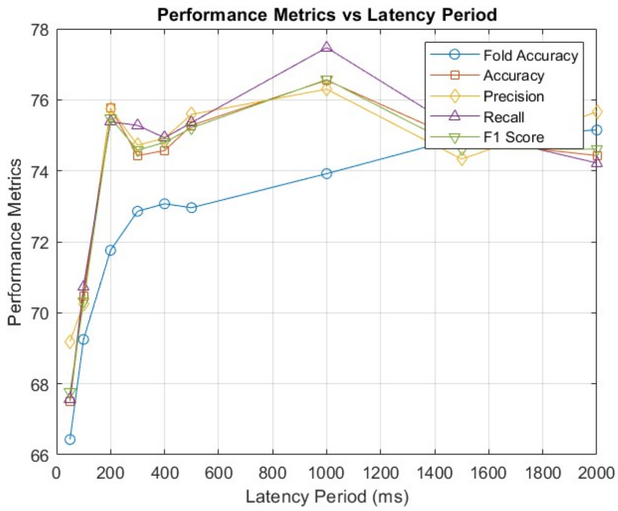

Fig. 4: Latency Period Analysis for different frames of UA-Speech

V. Summary And Conclusions

The study suggested utilizing LP residual-based MGDCC features to classify dysarthria severity-levels, using two established dysarthria speech corpora, namely, UA-Speech and TORGO. Through various experiments, we discovered important insights demonstrating the effectiveness of MGDCC features for classifying dysarthric severity. Incorporating phase-based features into a modified group delay function for dysarthric speech involves enhancing the analysis of speech signals by integrating phase information to improve recognition and intelligibility. The group delay function, which typically analyzes phase variations with frequency, is adjusted to better capture the unique spectral and phase distortions present in dysarthric speech.

A comparative analysis with baseline features, such as MFCC and MGDCC, revealed that LFCC features are superior for classifying dysarthria severity-levels. These findings were especially noteworthy because LFCC features exhibited performance at par with the advanced Whisper model features, highlighting their effectiveness in capturing dysarthria-specific characteristics.

Further analysis of the latency period illustrates how varying latency periods (50 ms to 2000 ms) affect performance metrics like accuracy, precision, recall, and F1 score. Shorter latencies may speed up decision-making but could reduce accuracy, while longer latencies often improve accuracy at the cost of increased processing time. The plot helps balance decision speed with model performance to find an optimal latency period.

One limitation of this work is the lack of direct comparison with current state-of-the-art deep learning techniques. This limitation is acknowledged due to constraints such as resource availability or differences in the scope and application of the study. While the primary focus has been on developing and validating a novel approach, future work should include comprehensive benchmarking against leading deep learning methods to more thoroughly assess the relative performance and generalizability of the proposed solution. Despite this, the results presented offer significant value, particularly in specialized scenarios where the state-of-the-art may not be directly applicable.

VI. Acknowledegements

The authors sincerely thank the Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, for sponsoring a consortium project titled 'BHASHINI', (Grant ID : 11(1)2022-HCC (TDIL)). They also thank the authorities of DA-IICT Gandhinagar for their kind support and cooperation in carrying out this research work.

References

[1] J. R. Duffy *et al.*, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Elsevier Health Sciences, 2012.

[2] L. P. Sahu and G. Pradhan, "Significance of filterbank structure for capturing dysarthric information through cepstral coefficients," IISc Banglore, India, 2022, pp. 1–5.

[3] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease," *The journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.

[4] H. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," vol. 1, May 2003, pp. I–68.

[5] S. Sajiha, K. Radha, D. Venkata Rao, N. Sneha, S. Gunnam, and D. Bavirisetti, "Automatic dysarthria detection and severity level assessment using CWT-layered cnn model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, Jun. 2024.

[6] Srivastava, "Fundamentals of linear prediction," 1999.

[7] B. Yegnanarayana, "Group delay spectrogram of speech signals without phase wrapping," *The Journal of the Acoustical Society of America(JASA)*, vol. 151, no. 3, pp. 2181–2191, 2022.

[8] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," *speech communication*, vol. 10, no. 3, pp. 209–221, 1991.

[9] H. Kim, M. Hasegawa-Johnson, A. Perlman, *et al.*, "Dysarthric speech database for universal access research.," in *Interspeech*, vol. 2008, 2008, pp. 1741–1744.

[10] 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.