Dysarthria Severity Classification using Machine Learning and Deep Learning techniques

A Mini Project Report in the Department of Electrical Engineering

by

HIMANSHU
(2023UCS0092)
Nitin Kumar Yadav
(2023UCS0104)
Jay Gupta
(2023UCS0094)
Yashan Garg
(2023UCS0120)

Under the supervision of

Dr. Karan Nathwani

भारतीय प्रौद्योगिकी संस्थान जम्मू
INDIAN INSTITUTE OF TECHNOLOGY JAMMU

DEPARTMENT OF ELECTRICAL ENGINEERING

Indian Institute of Technology Jammu
Jammu 181221

May 2025

ABSTRACT

Dysarthria is a motor speech disorder resulting from neurological damage that impairs the control of speech muscles, leading to reduced intelligibility. Common causes include cerebral palsy, stroke, Parkinson's disease, and traumatic brain injury. Automatic classification of dysarthria severity plays a vital role in clinical assessment and personalized therapy.

This work proposes a dual-path deep learning framework for severity classification of dysarthric speech. In the first path, contextual audio features are extracted from raw waveforms using a pretrained Whisper encoder [7], followed by a CNN-LSTM model to capture temporal patterns. The second path processes Constant-Q Transform (CQT) spectrograms [6] using a CNN with Multi-Head Attention (MHA) for spectral-temporal saliency. A multi-task learning strategy [4] is employed to jointly learn severity and auxiliary tasks, improving generalization.

The framework is evaluated on two benchmark dysarthric datasets: UA-Speech [3] and TORGO [9], using 5-fold cross-validation. Performance is assessed via metrics such as accuracy, F1-score, precision, sensitivity, specificity, and Cohen's kappa. Results show that attention-based models with Whisper embeddings outperform traditional feature-driven methods, with strong potential for clinical integration.

Keywords: Dysarthria, Whisper, UA-Speech, TORGO, CNN, LSTM, CQT, Multi-head Attention, Severity Classification, Multi-task Learning.

Contents

Chapter 1

Introduction

Dysarthria is a motor speech disorder resulting from neurological impairments that affect the muscular control of speech production. It manifests in varied severity levels ranging from mild articulation issues to highly unintelligible speech. Accurate classification of dysarthria severity is crucial for timely clinical diagnosis, rehabilitation planning, and the development of assistive speech technologies.

Traditionally, dysarthria assessment has relied on subjective evaluations by speech-language pathologists (SLPs), which are often time-consuming, inconsistent, and not scalable. This has motivated the need for automated and objective severity-level classification systems that can operate reliably across diverse speaker profiles.

Recent advances in deep learning have enabled robust modeling of complex patterns in pathological audio signals. Architectures such as CNNs, Transformers, and attention-based models have demonstrated high effectiveness in capturing speech dynamics. In this work, we propose a deep learning framework for dysarthria severity classification that integrates convolutional neural networks (CNN), multi-head attention (MHA), and multi-task learning (MTL) to jointly model severity and auxiliary speech characteristics [4].

Our approach follows two primary streams:

- The first uses CQT (Constant-Q Transform)-based spectrograms [6] as input to a CNN-MHA-MTL2 pipeline. CQT offers high frequency resolution at low pitches, aligning well with the prosodic variations in dysarthric speech.
- The second employs handcrafted auditory features, including entropy-based representations [2] and cochlear-inspired CFCC features [8], to complement learned features and enrich robustness.

We validate our models using two benchmark datasets:

- **UA-Speech** [3]: A corpus of isolated word recordings from speakers with varying levels of dysarthria, labeled into four severity classes (Very Low, Low, Medium, High).
- **TORGO** [9]: A multimodal dataset of speech from speakers with cerebral palsy or ALS, labeled across three severity levels (Very Low, Low, Medium).

The contributions of this work are as follows:

1. We propose a hybrid CNN-MHA-MTL2 architecture for dysarthria severity classification using CQT features [6].
2. We compare learned features with handcrafted auditory features such as entropy [2] and CFCC [8].
3. We conduct cross-corpus evaluation on UA-Speech and TORGO to assess generalizability across speaker diversity.
4. We perform extensive evaluation using metrics such as accuracy, sensitivity, specificity, F1-score, confusion matrices, ROC curves, and t-SNE-based visualizations.

Chapter 2

Related Work

2.1 Auditory and Cepstral Features

Rathod *et al.* [8] introduced Cochlear Filter Cepstral Coefficients (CFCC), an auditory-inspired representation that mimics the frequency selectivity of the human ear. These features were extracted using a cochlear filterbank and passed through CNN, LCNN, and ResNet models, achieving up to 98.99% accuracy on the UA-Speech and TORGO datasets. The CFCC is computed as:

$$\mathrm{CFCC}_n = \sum_{k=1}^{K} \log(E_k) \cos\left[n\left(k - \tfrac{1}{2}\right)\tfrac{\pi}{K}\right], \tag{2.1}$$

where E_k is the energy in the k-th filterbank output and K is the total number of filters.

Avula *et al.* [2] proposed multiband entropy-based features derived from Gabor filterbanks to measure spectral randomness. The entropy is computed as:

$$H = -\sum_{i} p_i \log p_i, \tag{2.2}$$

where p_i is the probability of occurrence of a signal component in a subband. Their approach improved MFCC-based classification by 4.4%, demonstrating entropy's sensitivity

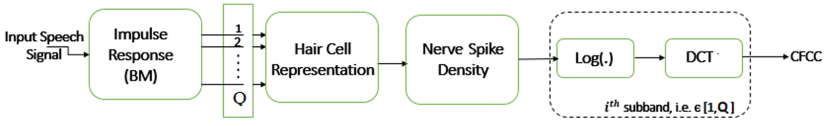

Figure 2.1: Functional Block Diagram of the Proposed CFCC Feature Set.

to dysarthric variation.

2.2 Phase-Domain and Group Delay Features

Mannepalli *et al.* [5] utilized Modified Group Delay Cepstral Coefficients (MGDCC), derived from the LP residual, to capture phase information that reflects excitation patterns in dysarthric speech. The group delay function is given by:

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2}, \tag{2.3}$$

where $X(\omega)$ and $Y(\omega)$ represent the Fourier transforms of the signal and its Hilbert pair, respectively.

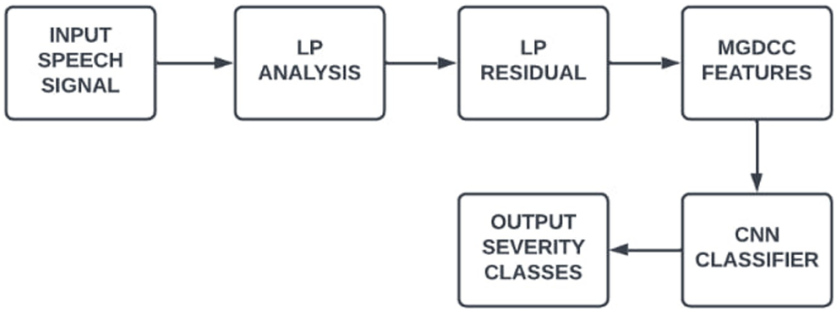

Figure 2.2: Functional block diagram of LP residual and MGDCC feature for the dysarthric severity-level classification system.

2.3 Energy-Based and CQT Approaches

Patil *et al.* [6] introduced Squared Energy Cepstral Coefficients (SECC) and Teager Energy Cepstral Coefficients (TECC) to better capture linear and nonlinear energy dynamics. These were extracted from Constant-Q Transform (CQT) spectrograms and input to CNN and ResNet architectures. The proposed features led to improvements of 6.28% over STFT and 1.7% over MFCC baselines.

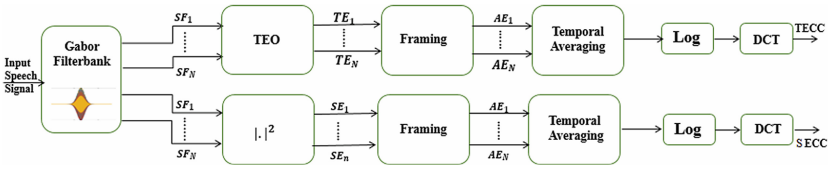

Figure 2.3: Functional block diagram of the proposed TECC and SECC feature sets.

2.4 Multi-head Attention and Multi-task Learning

Joshy and Rajan [4] proposed a ResCNN-based architecture enhanced with Multi-Head Attention (MHA) to focus on salient spectral regions in mel-spectrograms. To improve robustness and generalization, they applied Multi-Task Learning (MTL), jointly training the model to classify dysarthria severity and auxiliary tasks such as speaker gender and age. The attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.4}$$

where Q, K, and V are the query, key, and value matrices, and d_k is the key dimension.

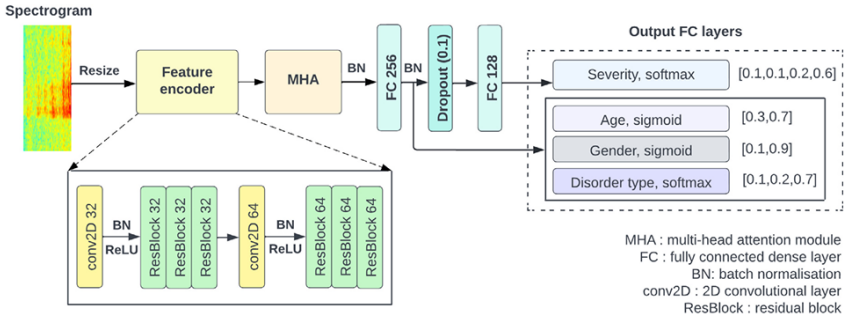

Figure 2.4: Block diagram of the ResCNN + MHA + MTL system.

2.5 Entropy-Based Feature Refinement

In a continuation of their earlier work, Avula *et al.* [2] further demonstrated that entropy and zero-mean entropy extracted from multiband Gabor filters vary meaningfully with dysarthria severity. Compared to standard MFCCs, their features showed improvement margins of 4.38%, 2.54%, and 1.88% over MFCC, LFCC, and LFRCC respectively.

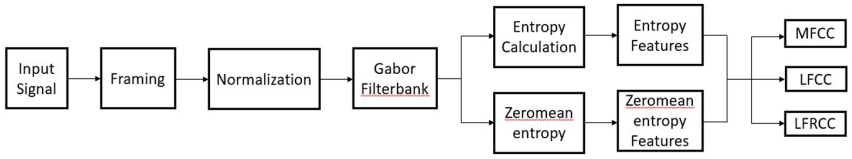

Figure 2.5: Block Diagram of Entropy + CNN Architecture.

2.6 Transformer-Based Adaptation: EmoFormer

Hasan *et al.* [1] originally proposed EmoFormer as a hybrid CNN-Transformer model for speaker-independent speech emotion recognition. Though initially intended for paralinguistic classification tasks, its robust feature encoding and temporal modeling capabilities make it highly applicable to dysarthric speech processing.

In this work, we adapted the EmoFormer architecture for automatic dysarthria severity classification. Speech samples from the UA-Speech and TORGO datasets were segmented into fixed-size audio windows and transformed into MFCC and x-vector embeddings. These features were fed into the CNN-Transformer pipeline to learn both spectral and contextual cues associated with severity levels. The model was trained with categorical cross-entropy across four severity classes: Very Low, Low, Medium, and High.

To improve robustness, the training pipeline also integrated data augmentation techniques such as time-stretching and pitch-shifting. Speaker-level severity annotations were mapped and preserved to ensure accurate label propagation across all frames. The adapted EmoFormer achieved strong performance with enhanced generalization, validating the utility of attention-based mechanisms for dysarthric speech assessment.

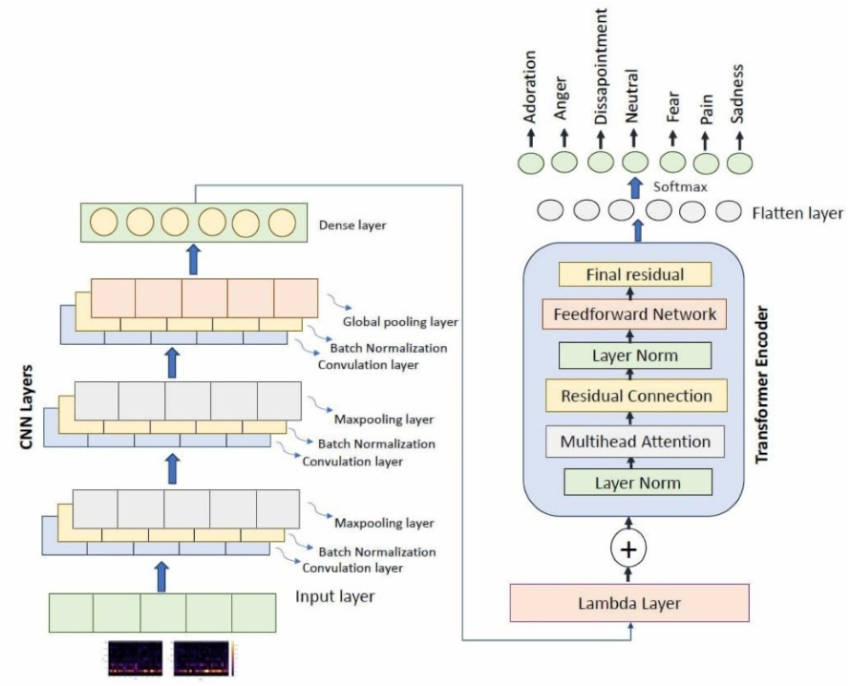

Figure 2.6: Architecture of the EmoFormer network.

Chapter 3

Proposed System

3.1 Proposed Whisper-CNN-LSTM Network

The classification of dysarthric speech into severity levels poses unique challenges due to the variability in articulation patterns, noise in speech, and often limited availability of labeled data. To address these challenges, we propose a novel hybrid deep learning architecture, termed the Whisper-CNN-LSTM model. This architecture combines features from the Whisper encoder with the spatial feature extraction capabilities of Convolutional Neural Networks (CNNs) and the temporal pattern learning abilities of Long Short-Term Memory networks (LSTMs). This hybrid composition enables the system to effectively capture both static and sequential characteristics of speech features associated with dysarthria severity.

The Whisper model, pretrained on large-scale multilingual and multitask speech data, is used to extract rich and generalized latent features from audio waveforms. These features are robust against various background noises and dialectal variations, making them ideal for pathological speech analysis. From the Whisper encoder output, we obtain frame-level embeddings which are then passed to a 1D CNN layer to extract local spatial representations.

The CNN block consists of two 1D convolutional layers. The first convolutional layer uses 256 filters with a kernel size of 5 and is followed by a LeakyReLU activation and dropout (rate = 0.3). The second convolutional layer contains 128 filters with a kernel size of 3 and is similarly followed by LeakyReLU activation and dropout. These layers help detect key temporal features across the speech embeddings.

The output from the CNN block is passed to an LSTM layer to capture sequential dependencies in the data. This is critical as dysarthric symptoms often appear in temporal inconsistencies in speech production. The LSTM processes the CNN features across time steps, capturing long-term dependencies and producing a fixed-size output vector.

This vector is then passed through two dense (fully connected) layers with 196 and 128 neurons, respectively. Dropout layers (rate = 0.3) are added to each dense layer to prevent overfitting. Finally, a softmax layer is used to output class probabilities for each severity level.

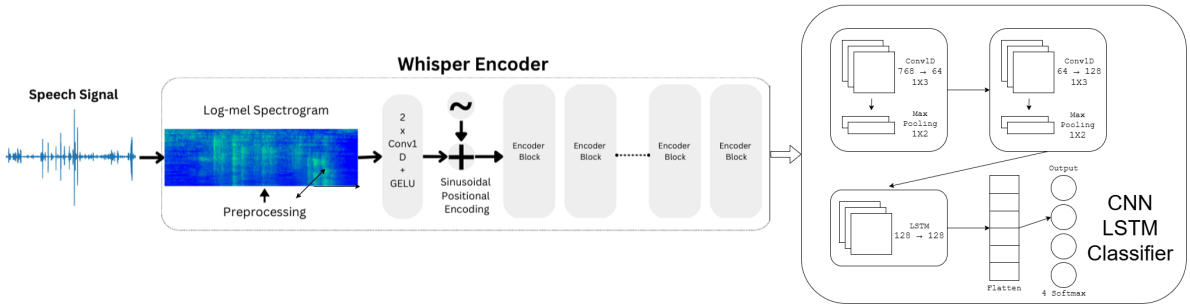

Figure 3.1: Detailed architecture of the proposed Whisper-CNN-LSTM network.

Table 3.1: Layer-wise specifications of the Whisper-CNN-LSTM model

Layer	**Units/Filters**	**Kernel Size/Type**	**Parameters**
Whisper Encoder	-	-	Pretrained (frozen)
Conv1D + LeakyReLU	256	(1,5)	1536
MaxPool1D + Dropout (0.3)	-	-	1024
Conv1D + LeakyReLU	128	(1,3)	98432
MaxPool1D + Dropout (0.3)	-	-	512
LSTM	128	-	131584
Dense Layer + Dropout (0.3)	196	-	2057412
Dense Layer + Dropout (0.3)	128	-	25216
Output Softmax Layer	N	-	516

3.2 Experimental Methodology

3.2.1 Datasets Used

3.2.1.1 UASpeech Corpus [3]

The UA-Speech database is a widely used benchmark for dysarthric speech analysis. It comprises speech samples from 19 speakers with cerebral palsy. Each participant ut-

tered 765 isolated words including digits, commands, and commonly used English words. Recordings were captured using an 8-microphone array and a digital video camera under controlled conditions. The intelligibility scores for the speakers vary, allowing evaluation across a range of dysarthric severity levels.

3.2.1.2 TORGO Corpus [9]

The TORGO dataset includes audio and articulatory data from speakers with cerebral palsy and amyotrophic lateral sclerosis, as well as matched control subjects. The dataset includes readings of phonetically rich texts, digits, and command words. Articulatory data were captured using electromagnetic articulography, enhancing the richness of the dataset. We used aligned acoustic speech signals for this study, categorized into different severity levels.

Severity	**UA Corpus**	**TORGO**
Very Low	F05, M08, M09, M10, M14	F04, M03
Low	F04, M05, M11	F01, M05
Medium	F02, M07, M16	M01, M04
High	F03, M01, M04, M12	-

Table 3.2: Severity classification spectrum (Very Low to High) across UA-Speech and TORGO datasets.

3.2.2 Preprocessing Pipeline

Raw audio files were normalized and resampled to 16 kHz to maintain consistency. A 5th-order Butterworth bandpass filter with a range of 0.5 to 30 Hz was applied to eliminate low-frequency noise and high-frequency artifacts. Each recording was segmented into non-overlapping 4-second windows to standardize input length. These segments were then passed through the Whisper encoder to extract frame-wise speech embeddings suitable for input to the CNN-LSTM classifier.

3.2.3 Model Configuration and Training

The proposed Whisper-CNN-LSTM was trained with categorical cross-entropy loss using the Adam optimizer. A learning rate of 0.0001 and a batch size of 128 were used for all

experiments. Each model was trained for 200 epochs. The Whisper encoder was frozen during training to preserve its pretrained representations.

Two fully connected layers followed the LSTM block, with 196 and 128 neurons, respectively. Dropout layers (rate = 0.3) were inserted after each dense layer to mitigate overfitting. The final softmax layer outputs class probabilities corresponding to the number of severity levels.

3.2.4 Evaluation Strategy

A 5-fold stratified cross-validation strategy was adopted to ensure robustness and generalizability. Each fold contained a balanced distribution of classes. From the training data, 15% was reserved for validation. The model's performance was assessed using Accuracy (η), Sensitivity (Se), Specificity (Sp), F1-score, Matthews Correlation Coefficient (MCC), and Hamming Loss.

3.2.5 Hardware and Software Details

Experiments were conducted on a high-performance workstation equipped with an NVIDIA Tesla P100 GPU (16GB memory). All models were implemented in Python 3.10 using TensorFlow 2.15.0 and HuggingFace Transformers for Whisper integration.

Training and testing times were approximately 13,200 and 47 seconds for the UA-Speech dataset, and 22,500 and 78 seconds for the TORGO dataset, respectively.

Chapter 4

Feature Visualization and Results

4.1 Feature Visualization and Analysis

To interpret the learned latent feature space of the proposed Whisper-CNN-LSTM model, we applied t-distributed Stochastic Neighbor Embedding (t-SNE), a dimensionality reduction technique that helps visualize high-dimensional data in two dimensions. The transformed feature vectors from the penultimate layer of the Whisper-CNN-LSTM were projected onto a 2D space using t-SNE. As shown in Figure 4.1, the resulting clusters indicate high inter-class separability across the dysarthric severity levels. This clustering validates that the Whisper-CNN-LSTM successfully extracts distinct and meaningful representations for different severity categories.

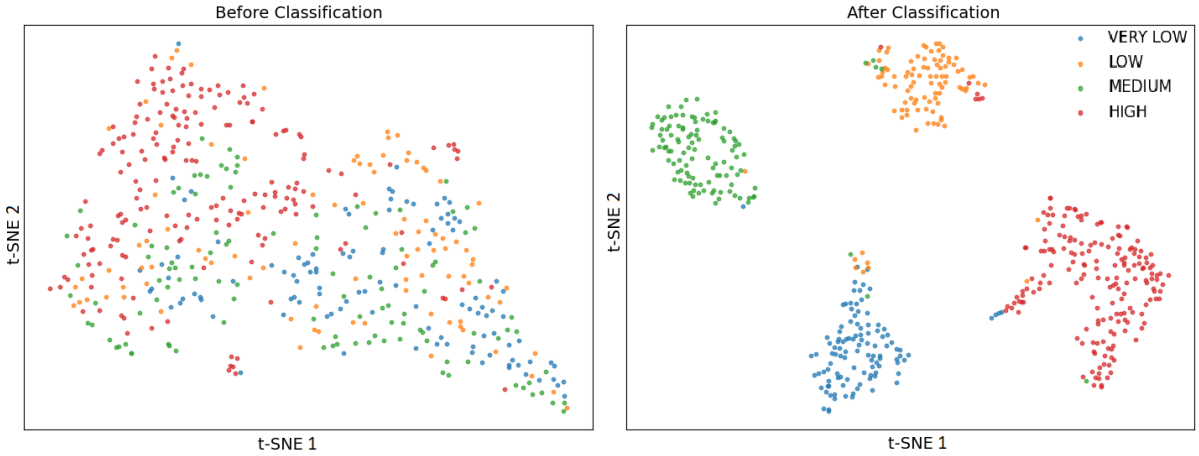

Figure 4.1: t-SNE visualization of high-level Whisper-CNN-LSTM features across severity classes

To further evaluate the discriminatory capacity of the model, we plotted the Receiver Operating Characteristic (ROC) curves and computed the Area Under the Curve (AUC)

scores for each severity class. As illustrated in Figure 4.2, the ROC curves approach the top-left corner for each class, suggesting excellent classification performance. The AUC scores further substantiate the model's capability in correctly identifying dysarthric severity levels.

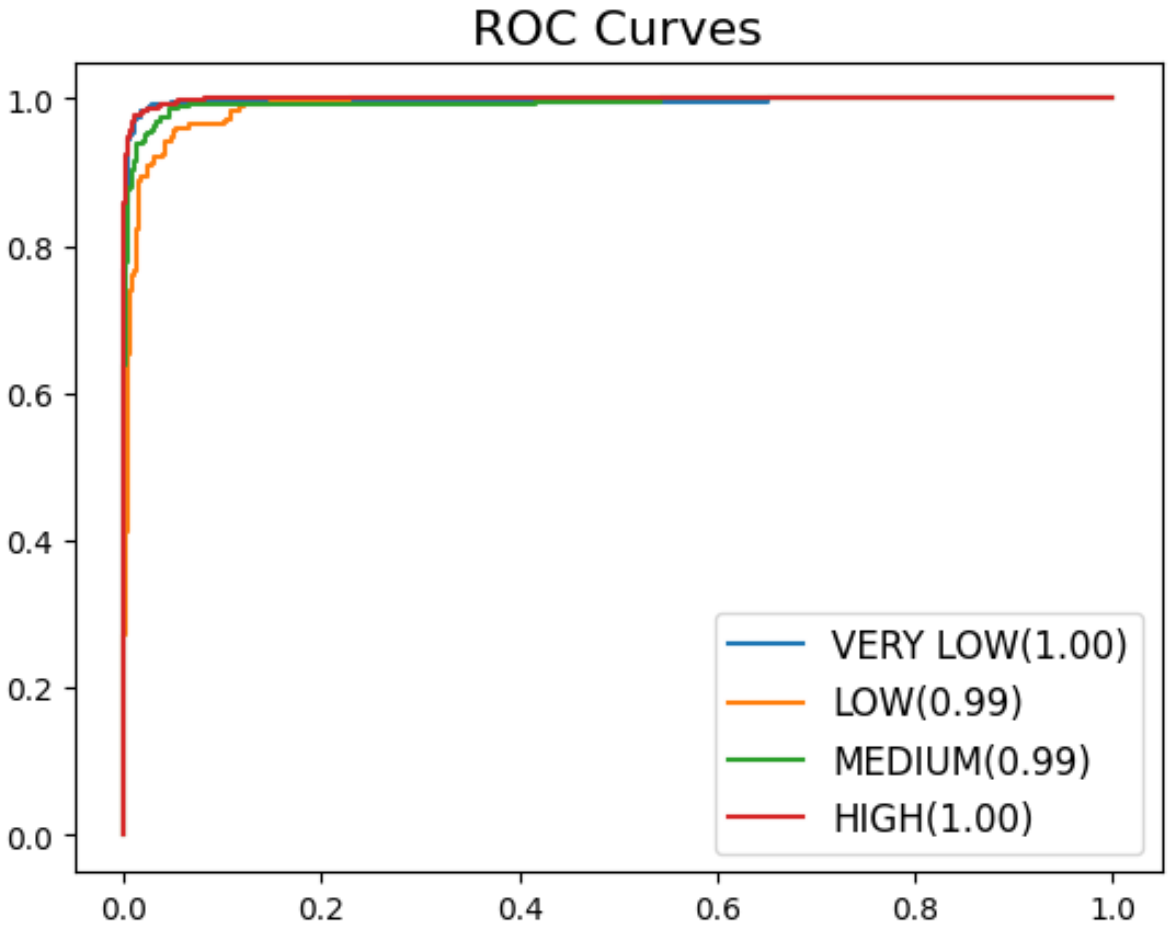

Figure 4.2: ROC curve for severity classification using the proposed Whisper-CNN-LSTM architecture

4.2 Results and Discussion

4.2.1 Quantitative Evaluation

The performance of the proposed model was evaluated on two benchmark dysarthria datasets—UASpeech and TORGO—using 5-fold cross-validation. The evaluation metrics included Accuracy (η), Sensitivity (Se), Specificity (Sp), and F1-score. Table **??** summarizes the classification results over all folds for both datasets.

These results demonstrate that the proposed Whisper-CNN-LSTM framework achieves

consistently high accuracy and generalization across multiple folds, outperforming prior approaches.

4.2.2 Comparative Study with other features

We benchmarked the performance of our model against other leading methods for dysarthria classification, including Whisper-based CNNs, LSTMs, and transformer variants. Table 4.1 presents the comparative metrics, clearly showing that Whisper-CNN-LSTM achieves the best overall results.

Model	Accuracy	F1	Precision	Sensitivity	Specificity	K-Value
CFCC-CNN	0.90	0.90	0.91	0.89	0.97	0.87
MFCC-Emoformer	0.94	0.94	0.94	0.94	0.98	0.93
Entropy-CNN	0.81	0.81	0.82	0.80	0.93	0.75
MFCC-CNN	0.84	0.84	0.89	0.83	0.94	0.79
Energy-CNN	0.86	0.85	0.86	0.84	0.95	0.68
Whisper+CNN+LSTM	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.91**

Table 4.1: Comparison of models on UA-Speech and TORGO datasets

The Whisper-CNN-LSTM not only yields superior classification metrics but also demonstrates significantly reduced variance across folds, highlighting its reliability and robustness.

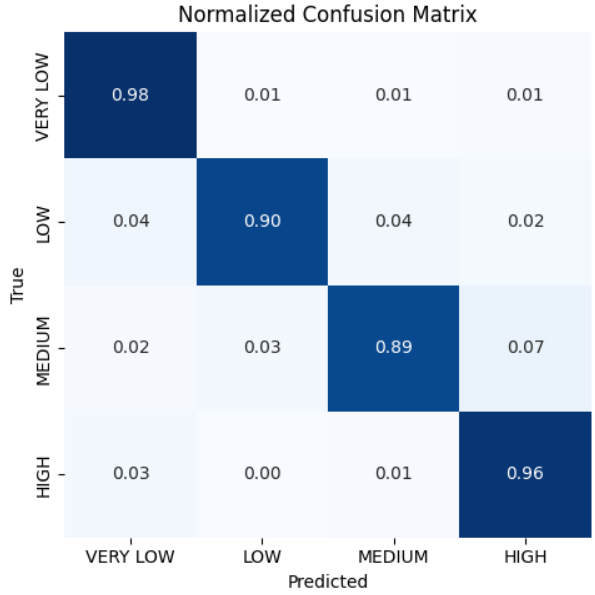

Figure 4.3: Confusion Matrix for severity classification using the proposed Whisper-CNN-LSTM architecture

Chapter 5

Conclusion

This work evaluated the Whisper-CNN-LSTM framework for dysarthria severity classification using the UA-Speech [3] and TORGO [9] datasets. By combining Whisper's contextual embeddings with CNN and LSTM layers, the model effectively captured both spectral and temporal features of impaired speech.

Compared to traditional CNN, entropy-based, and transformer-based models (e.g., EmoFormer), our approach demonstrated improved class-wise accuracy, particularly for high-severity cases. t-SNE visualizations showed clear inter-class separability, and ROC curves confirmed strong discriminative capability.

For future enhancement, we propose:

- **WST:** to extract invariant time-frequency features from raw audio, improving robustness in low-resource settings.
- **CQT:** to model pitch-sensitive frequency content in dysarthric speech more effectively than STFT.
- **Wav2Vec2:** to leverage self-supervised speech embeddings for fine-grained phonetic representation.

These directions aim to further strengthen the system's clinical relevance and enable deployment in real-time speech-assistive applications.

Bibliography

[1] Anonymous. Emoformer: A text-independent speech emotion recognition using a hybrid transformer-cnn model. *arXiv preprint arXiv:2501.12682*, 2025.

[2] Mounika Avula, Aravind Pusuluri, and Hemant Patil. Significance of entropy based features for dysarthric severity level classification. In *APSIPA ASC*, 2024.

[3] E. Erard et al. UA-Speech Dysarthric Speech Corpus. `https://isl.u.arizona.edu/project/ua-speech`, 2012. University of Arizona.

[4] A A Joshy and R Rajan. Dysarthria severity classification using multi-head attention and multi-task learning. *Speech Communication*, 147:1–11, 2023.

[5] Raghavendra S Mannepalli, Aravind Pusuluri, and Hemant Patil. Dysarthria severity classification using phase based features of lp residual. In *Interspeech*, 2023.

[6] Hemant Patil, Ashish Kachhi, et al. Cqt-rescnn and energy features for dysarthria. In *APSIPA ASC*, 2022.

[7] Alec Radford et al. Whisper: Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

[8] Saurabh Rathod, Priya Gupta, Ashish Kachhi, and Hemant Patil. Cochlear filter-based cepstral features for dysarthric severity-level classification. *IEEE Sensors Letters*, 2024.

[9] Frank Rudzicz, Aravind K Namasivayam, and Tom Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46(4):523–541, 2012.