Mini Project Presentation on

Dysarthria Severity Classification using Machine Learning and Deep Learning Techniques

Presented By:

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

Supervisor:

Dr. Karan Nathwani, Associate Professor

Department of Electrical Engineering
Indian Institute of Technology Jammu

May, 2025

Flow of Presentation

1. Introduction
2. Objectives
3. Literature Review
4. Proposed Method
5. Experimental Methodology
6. Results and Discussion
7. Conclusion
8. Future Work

Introduction

- **What is Dysarthria?**
  A motor speech disorder resulting from impaired muscular control due to neurological damage.
- **Causes:**
  Stroke, brain injury, Parkinson's disease, ALS, or cerebral palsy.
- **Reason:**
  Disruption in the communication between the brain and speech muscles affects articulation, voice, and breath control.
- **Advantages of Automated Detection:**
  Faster diagnosis, consistent evaluation, early intervention, and scalable screening.
- **Disadvantages and Challenges:**
  Data scarcity, variability in speech patterns, need for personalization.
- **Applications:**
  Clinical decision support, telehealth, rehabilitation monitoring, and assistive communication.

Type of Dysarthria

Type	**Localization**
Flaccid	Lower motor neuron
Spastic	Bilateral motor neuron, Unilateral upper motor neuron
Ataxic	Cerebellum
Hypokinetic	Extrapyramidal
Hyperkinetic	Extrapyramidal
Spastic and flaccid	Upper and lower motor neuron

Table 1: Classification of Dysarthria

[0] *Source: Duffy, J.R., 1995*

Severity Levels and Clinical Relevance

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

IIT JAMMU

[0] https://www.isca-archive.org/interspeech$_2$023/$rathod23_interspeech.pdf$

Objectives

- To develop an automatic classification system for classifying dysarthric speech into severity levels.
- To compare handcrafted auditory features with learned representations.
- To validate the different deep learning models on UA-Speech and TORGO datasets.
- To visualize learned features and assess model robustness across speaker variations.
- To develop a hybrid deep learning architecture using Whisper encoder, CNN, and LSTM.

Various studies have addressed dysarthria severity classification using handcrafted, statistical, and deep features:

- **Entropy-based methods** [1] extracted multiband entropy and zero-mean entropy using Gabor filterbanks, improving classification accuracy by capturing randomness in dysarthric speech.
- **Phase-based features** [4] used Modified Group Delay Cepstral Coefficients (MGDCC) from LP residuals, effectively leveraging glottal excitation and phase characteristics for severity detection.
- **Multi-task learning with attention** [3] employed ResCNN with Multi-Head Attention (MHA) and auxiliary tasks (age, gender, disorder type) to improve robustness and inter-class discrimination on UA-Speech.
- **CQTCNN+ResCNN** [5] applied Constant-Q spectrograms to capture time-frequency features more precisely than STFT or MFCC, achieving notable gains with CNN-based classifiers.
- **Cochlear Filter-Based Features** [6] introduced CFCCs inspired by human auditory modeling, which showed superior performance on both UA-Speech and TORGO datasets when evaluated with CNN, LCNN, and ResNet.
- **EmoFormer** [2] combined CNN and Transformer blocks for text-independent speech emotion recognition, showing the advantage of attention-based models for generalizing across speakers and tasks.

Paper	Features	Model	Dataset	Accuracy
Joshy et al.[3]	Mel Spectrogram	ResCNN + MHA + MTL	UA-Speech	↑11.5%
Mannepalli et al. [4]	MGDF + LP Residual	Stratified CNN	UA-Speech, TORGO	High
Avula et al.[1]	Entropy + MFCC	CNN	UA-Speech	↑4.38% over MFCC
Rathod et al.[6]	CFCC	CNN, LCNN, ResNet	UA-Speech, TORGO	Up to 98.99%
Hasan et al.[2]	MFCC, X-Vector	CNN + Transformer	EARS	90% (5 emotions)

Table 3: Illustration of State-of-the-Art methods

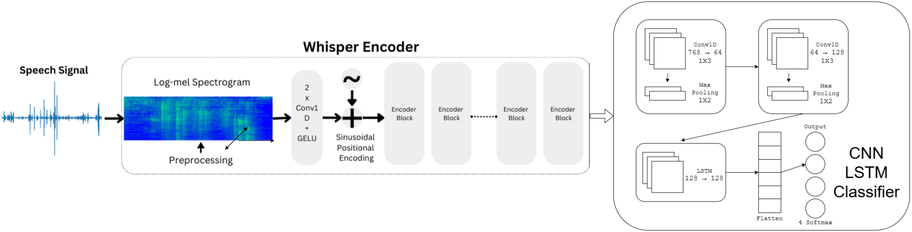

Figure 1: An outline of the proposed concept

Whisper + CNN + LSTM Architecture

- **Input:**
  - Raw audio waveform sampled at **44.1 kHz**.
  - Normalized and preprocessed for Whisper encoder.
- **Whisper Encoder:**
  - Outputs high-dimensional latent features (80-d log-Mel-like vectors).
  - Trained on 680,000 hours of speech; highly generalized representations.
- **CNN Layers:**
  - 2D convolutions extract local articulatory distortions.
  - Includes BatchNorm, ReLU, and MaxPooling.
- **LSTM Layers:**
  - Bidirectional LSTM (2 layers, 128 units) to capture speech dynamics.
  - Helps model changes in speech fluency, rhythm, and energy.
- **Output Layer:**
  - Dense layer with Softmax (4 neurons for 4 severity levels).
  - Outputs class probabilities.

IIT JAMMU

Experimental Methodology

- **Datasets Used:**
  - **UA-Speech:** Contains recordings from 15 dysarthric and 13 control speakers with labeled severity.
  - **TORGO:** Includes dysarthric and control speech from individuals with cerebral palsy and ALS.
- **Preprocessing:**
  - Audio downsampled to **44.1 kHz**, normalized, and optionally denoised.
  - Silences trimmed using energy-based VAD.
- **Feature Extraction:**
  - Whisper encoder produces 80-dimensional embeddings per time frame.
  - Embeddings passed to CNN-LSTM model for severity classification.
- **Training Setup:**
  - Optimizer: Adam, learning rate = 1e-4
  - Loss Function: Categorical Cross-Entropy
  - Batch Size: 32, Epochs: 50
  - 5-Fold Stratified Cross-Validation
- **Evaluation Metrics:**
  - Accuracy, Precision, Recall, F1-Score
  - Confusion Matrix analysis for misclassification insights

Performance Metrics:

- The following metrics were used to evaluate classification performance:

$$\text{Accuracy (\%)} = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{6}$$

where p_o is the observed accuracy, and p_e is the expected agreement by chance.

Severity Class	Accuracy (%)
Very Low	98.67
Low	89.39
Medium	90.64
High	95.52

Table 4: Whisper-CNN-LSTM: Class-wise accuracy

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 5: Comparison of models on UA-Speech and TORGO datasets

Confusion Matrix for Severity-Level Classification

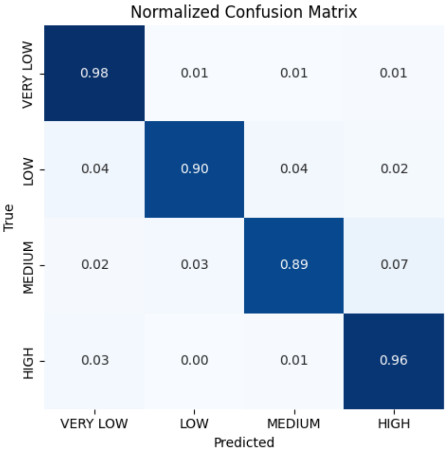

Figure 2: Confusion matrix showing true vs predicted severity levels for dysarthric speech

IIT JAMMU

t-SNE Visualization of Learned Features

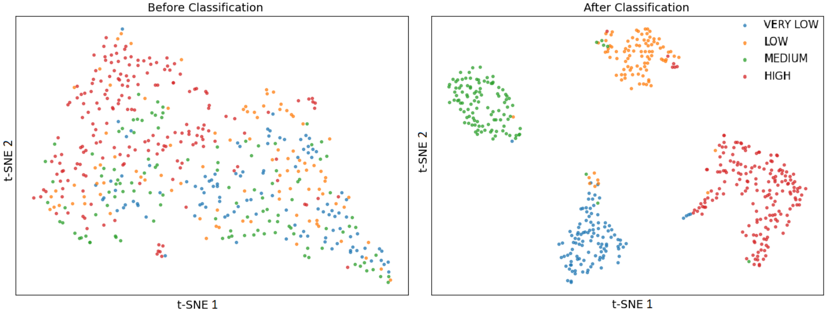

Figure 3: Visualization of Feature Embeddings for Dysarthria Severity Level Classification using t-SNE.

ROC Curves for Severity-Level Classification

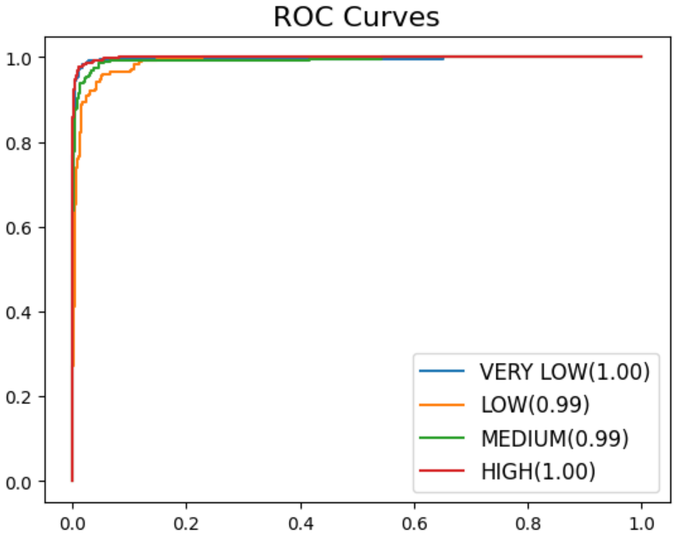

Figure 4: One-vs-Rest ROC curves for each dysarthria severity class.

IIT JAMMU

Conclusion

- We propose a hybrid deep learning model Whisper encoder + CNN + LSTM for dysarthria severity classification.
- The model effectively classifies speech into four severity levels: Very Low, Low, Medium, and High.
- Whisper embeddings provided robust and noise-invariant features from raw speech.
- CNN extracted spatial patterns and LSTM captured temporal dependencies in speech.
- Achieved an overall accuracy of 94% with high F1-scores across all classes.
- Experimental results on UA-Speech and TORGO datasets show improved performance over traditional MFCC-based baselines.

IIT JAMMU

Future Work

- Extend the proposed model to handle **multilingual dysarthric speech** by leveraging Whisper's large-scale multilingual pretraining.
- Evaluate model performance under **real-world noise conditions** and channel variability to ensure practical robustness.
- Explore **Transformer-based architectures**, such as EmoFormer and Wav2Vec2.0, for improved temporal modeling and speaker-independent generalization.
- Integrate **Wavelet Scattering Transform (WST)** as a fixed deep feature extractor to enhance stability and reduce reliance on large training data.
- Investigate **Constant-Q Transform (CQT)** as an alternative to STFT or Mel spectrograms for better capturing pitch and harmonic structure in dysarthric speech.
- Apply the model on **clinical-grade datasets** and explore deployment in real-time **assistive tools for speech-language pathologists (SLPs)**.
- Study **domain adaptation techniques** for cross-device, cross-dataset, and speaker-independent deployment.

[1] Mounika Avula, Aravind Pusuluri, and Hemant Patil. "Significance of Entropy Based Features For Dysarthric Severity Level Classification". In: *APSIPA ASC*. 2024 (cit. on pp. 7, 8).

[2] Rashedul Hasan. "EmoFormer: A Text-Independent Speech Emotion Recognition using a Hybrid Transformer-CNN model". In: *arXiv preprint arXiv:2501.12682* (2025) (cit. on pp. 7, 8).

[3] A A Joshy and R Rajan. "Dysarthria Severity Classification using Multi-head Attention and Multi-task Learning". In: *Speech Communication* 147 (2023), pp. 1–11 (cit. on pp. 7, 8).

[4] Raghavendra S Mannepalli, Aravind Pusuluri, and Hemant Patil. "Dysarthria Severity Classification Using Phase Based Features of LP Residual". In: *Interspeech*. 2023 (cit. on pp. 7, 8).

[5] Hemant Patil, Ashish Kachhi, et al. "CQT-ResCNN and Energy Features for Dysarthria". In: *APSIPA ASC*. 2022 (cit. on p. 7).

[6] Saurabh Rathod et al. "Cochlear Filter-Based Cepstral Features for Dysarthric Severity-Level Classification". In: *IEEE Sensors Letters* (2024) (cit. on pp. 7, 8).

Thank You