

A Multilingual Framework for Dysarthria: Detection, Severity Classification, Speech-to-Text, and Clean Speech Generation

Ananya Raghu^{1,*}, Anisha Raghu^{1,*}, Nithika Vivek^{1,*}, Sofie Budman^{1,*}, Omar Mansour^{1,2}

¹Beaver Works Summer Institute, Massachusetts Institute of Technology

²Columbia University

Emails: ananyaraghu10@gmail.com, anisharaghu10@gmail.com

nithika.vivek@gmail.com, sofiebudman@gmail.com, omar.mansour@columbia.edu

Abstract—Dysarthria is a motor speech disorder that results in slow and often incomprehensible speech. Speech intelligibility significantly impacts communication, leading to barriers in social interactions. Dysarthria is often a characteristic of neurological diseases including Parkinson’s and ALS, yet current tools lack generalizability across languages and levels of severity. In this study, we present a unified AI-based multilingual framework that addresses six key components: (1) binary dysarthria detection, (2) severity classification, (3) clean speech generation, (4) speech-to-text conversion, (5) emotion detection, and (6) voice cloning. We analyze datasets in English, Russian, and German, using spectrogram-based visualizations and acoustic feature extraction to inform model training. Our binary detection model achieved 97% accuracy across all three languages, demonstrating strong generalization across languages. The severity classification model also reached 97% test accuracy, with interpretable results showing model attention focused on lower harmonics. Our translation pipeline, trained on paired Russian dysarthric and clean speech, reconstructed intelligible outputs with low training (0.03) and test (0.06) L1 losses. Given the limited availability of English dysarthric-clean pairs, we finetuned the Russian model on English data and achieved improved losses of 0.02 (train) and 0.03 (test), highlighting the promise of cross-lingual transfer learning for low-resource settings. Our speech-to-text pipeline achieved a Word Error Rate of 0.1367 after three epochs, indicating accurate transcription on dysarthric speech and enabling downstream emotion recognition and voice cloning from transcribed speech. Overall, the results and products of this study can be used to diagnose dysarthria and improve communication and understanding for patients across different languages.

Index Terms—dysarthria, signal processing, machine learning, automatic speech recognition, speech synthesis, voice cloning

I. INTRODUCTION

A. Problem Statement

Dysarthria is a motor speech disorder and a common symptom of neurological conditions such as ALS, Parkinson’s disease, stroke, and cerebral palsy [1]. It arises when the nervous system damage impairs the muscles involved in speaking, leading to slurred, slow speech that is difficult to understand [2]. While not a disease itself, dysarthria significantly impairs communication and quality of life, frequently leading to social isolation, misdiagnosis, or reduced access to care. Studies report that dysarthria occurs in up to 60% of stroke patients and affects as many as 90% of individuals with Parkinson’s

disease [3]. Despite its prevalence, Dysarthria is often under-recognized, particularly in its milder forms or in multilingual populations [4]. A recent study demonstrated that a listener’s native language significantly influences their perceptual ratings of dysarthria, particularly for articulatory and rhythmic characteristics [5]. This highlights a fundamental limitation in human-based assessment, as a clinician’s ability to accurately perceive and rate a speaker’s dysarthria can be compromised when they are not a native speaker of the language. This suggests that current diagnostic methods that rely on subjective evaluations by speech-language pathologists are constrained by language familiarity and clinical access. Furthermore, they are prone to human error and bias, which can delay proper treatment, especially for early stage dysarthria [6], [7].

In contrast, machine learning models trained on diverse, labeled datasets offer an objective alternative to human assessments. By extracting language-agnostic acoustic features and learning features across speech samples, ML can reduce diagnostic bias and enable more consistent screening. This makes machine learning based tools especially promising for accessible dysarthria detection across diverse healthcare settings.

B. Prior Machine Learning Approaches

1) *Prior work: Dysarthria Detection:* Recent advances in machine learning have led to the development of models capable of detecting dysarthria using acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectrograms, or prosodic cues. Prior work has focused largely on binary classification, distinguishing dysarthric from healthy speech, using convolutional or recurrent neural networks trained on datasets like TORGO and UA-Speech [8], [9]. However, these models are typically trained and evaluated on a single language, limiting their clinical applicability across multilingual populations.

2) *Prior Work: Severity Classification:* While recent approaches [10] to dysarthria severity classification have received high performance using neural networks, they are often black boxes, not explaining the reasoning behind model classification. Furthermore, features extracted for the model, including embeddings from wav2vec2 [10], do not offer insight on which acoustic characteristics of the slurred speech distinguish it from clean or less severe dysarthric speech. An interpretable

* Equal contribution.

model can thereby give insight as to which features of the speech are altered in dysarthria, helping with speech therapy and increasing patient understanding.

3) *Prior work: Speech Synthesis:* Prior research has explored several generative voice conversion and augmentation approaches for translating dysarthric speech to regular speech. The CycleGAN-VC model architecture was applied to Korean dysarthric speech (18700) utterances and healthy controls reducing Word-Error-Rate by 33.4% [11]. However, CycleGANs often produce artifacts especially in highly impaired speech and pixel level consistency can potentially be problematic and cause unrealistic images [12]. The DVC 3.1 system, which combines data augmentation with a StarGAN-VC backbone [13] improved both ASR word recognition and listener ratings. Still, the quality of generated speech heavily depends on the synthetic data distribution and can degrade for highly variable input. More recently, diffusion-based voice conversion with Fuzzy Expectation Maximization (FEM) [14] improved intelligibility and accuracy using soft clustering, but diffusion models are typically slow to sample.

4) *Prior work: Speech to Text, Emotion, and Voice Cloning:* Prior work has explored speech to text pipelines on patients with dysarthria, marking the first step towards increased patient understanding [15]. However, these methods have stopped at the transcription stage and have not progressed to sentiment classification or synthetic speech generation. In order to improve communication and understanding of dysarthric patients, it is crucial to provide this population with a way to communicate using acoustic features of their voice prior to their dysarthric diagnosis, enabling better understanding of sentiment and needs.

II. METHODS

1) *Datasets:* We utilized four primary datasets in this study. The TORGO dataset [16] provides paired audio samples and textual prompts from individuals with and without dysarthria, supporting analysis of articulatory impairments in English Speech. From this corpus, we accessed a subset of approximately 2,000 audio files (500 each for female non-dysarthric, female dysarthric, male dysarthric, and male non-dysarthric speakers) made available on Kaggle [17], and paired them with textual prompts sourced from a separate Kaggle dataset [18]. The UA Dysarthria dataset [19] was used for severity classification and includes 11,436 speech spectrograms labeled across 4 severity levels: very low, low, medium, high. For Russian-language data, the Hyperkinetic Dysarthria Speech dataset [20] was utilized, providing 2000 samples from both Dysarthric and non-dysarthric patients reciting the same phrase, along with corresponding prompts. Additionally, the Dysarthric German dataset [21] contributed 1,272 samples of Dysarthric German speech for crosslingual prediction.

2) *Initial Feature Extraction:* To gain initial insight into speech patterns associated with dysarthria, we visualized spectrograms across gender and condition (dysarthric vs non-dysarthric). In the highlighted regions shown in Figure 1, we observe that dysarthric speech tends to exhibit prolonged

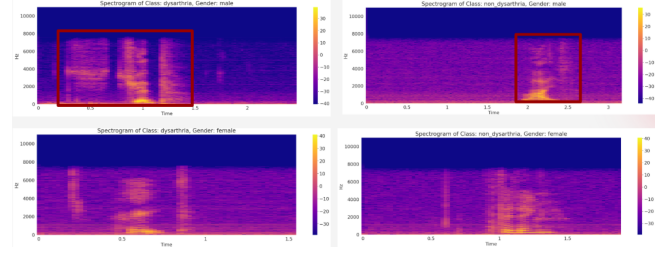


Fig. 1. Spectrogram visualizations of dysarthric and non-dysarthric speech across gender

low-frequency spectral bands and reduced clarity, indicative of slurred articulation and irregular pacing. In contrast, non-dysarthric speech shows more distinct high-frequency bursts and cleaner articulation boundaries.

A. Dysarthria Detection Task

Building on these qualitative differences, we extracted quantitative features for classification using Mel-Frequency Cepstral Coefficients (MFCCs), a widely used representation in speech processing. The pipeline begins with a Fast Fourier Transform to convert the raw audio into its frequency spectrum, followed by a logarithmic amplitude scaling that mimics human loudness perception. Mel scaling is then applied to emphasize perceptually relevant frequency bands. Finally, a Discrete Cosine Transform reduces dimensionality while preserving key spectral features. The resulting MFCCs capture articulatory and phonatory characteristics that are especially relevant for identifying dysarthric patterns.

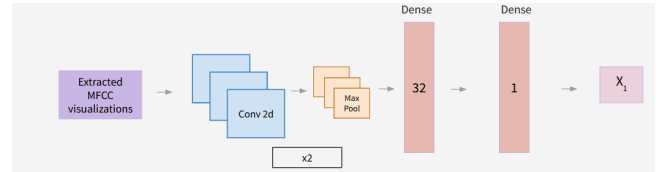


Fig. 2. Dysarthria Classification Model Architecture

To perform dysarthria detection from MFCC inputs, we adapted a simple 2D Convolutional Neural Network (CNN) architecture based on a publicly available Kaggle notebook. The model takes MFCC visualizations as input and passes them through two convolutional layers with max-pooling, followed by a dense layer with 32 units and a final output layer for binary classification. This architecture, as shown in Figure 2, is effective at capturing time-frequency patterns relevant to dysarthric speech for our multilingual classification experiments. The model was trained for 50 epochs after which the training and validation loss converged.

B. Severity Classification Task

The second stage of our project involved classifying each dysarthric patient's severity level as an indicator of how far along they are in their progression towards more serious diseases like ALS and Parkinson's. Accurate severity diagnosis

can help a patient tailor their healthcare plan and treatment accordingly to better suit their needs [22].

11436 spectrogram images of dysarthric audio were downloaded from Kaggle, resized to 128x128, converted to an array, and normalized. Class labels belonging to high severity (3036/11426), medium severity (2295/11426), low severity (2280/11426), and very low severity (3825/11426) were one-hot-encoded. The dataset split into train, test, and validation (70-20-10).

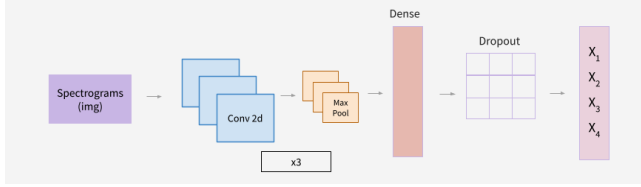


Fig. 3. Severity Classification Model Architecture

A sequential model was defined with model architecture shown in Figure 3. Three 2D convolutional layers were used to extract important features from the model, each of which were followed by a Max Pooling layer for dimensionality reduction. A dropout of 0.5 was specified to reduce overfitting and increase generalizability. The model was trained for 10 epochs after which the training and validation loss converged.

C. Multi-lingual clean speech synthesis

The third component of the proposed framework is a two-stage pipeline to translate dysarthric speech into normal speech. Translating dysarthric speech into more understandable forms is crucial because reduced intelligibility severely limits a dysarthric patient's ability to engage in daily conversations [23].

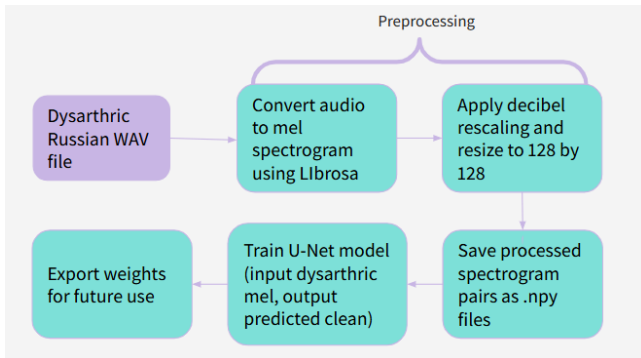


Fig. 4. Stage 1: Dysarthric Speech to Normal Speech (Russian)

1) Stage 1: Russian Dysarthric Speech to Normal Speech: In the first stage of our pipeline, we trained a U-Net model to map dysarthric speech to clean speech using Russian speech. As outlined in Figure 4, raw .wav files were converted into mel spectrograms using Librosa, then rescaled in decibels and resized to a uniform 128x128 image for consistent model input. These paired spectrograms were saved as .npy files for training. The U-Net model was trained for 300 epochs, with

a learning rate of 1e-4, to output a clean spectrogram from a dysarthric spectrogram, and the learned weights were exported. This step allows the model to learn structural transformations between distorted and healthy speech, that generalize across languages due to shared characteristics.

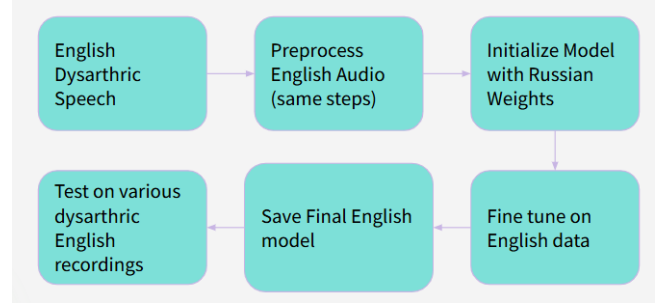


Fig. 5. Phase 2: Dysarthric Speech to Normal Speech (English)

2) Stage 2: Adapting Russian Model to English Dysarthric Speech: Although the Torgo database contains substantial English dysarthric speech, we encountered a key challenge: very few clean-dysarthric pairs were spoken with the same text, which is necessary for paired spectrogram training. To construct a usable dataset, we manually filtered for matched male and female speakers saying the same sentences. After preprocessing, we identified only 190 matched female and 37 matched male samples. These were processed into mel spectrograms using the same pipeline as in Stage 1. To address the limitations of training from scratch, we leveraged our U-Net model trained on the larger Russian dataset which had already learned to correct dysarthric distortions and followed the steps shown in Figure 5. We processed English dysarthric audio using the same preprocessing steps outlined in Phase 1, then initialized the model with Russian weights. We then fine tuned the model using the small available English dataset for 300 epochs and observed an improved performance over models trained from scratch. This cross-lingual transfer approach demonstrates how transfer learning can be used to compensate for scarcity of data and can be potentially applied to low-resource languages.

D. Automatic Speech Recognition, Emotion Classification, and Voice Cloning

The overall pipeline for automatic speech recognition, emotion classification, and voice cloning is shown in Figure 6. We start by converting the audio to text, then use the text to classify emotion and perform voice cloning.

1) Speech to Text: The fourth component of our pipeline is a speech-to-text converter for dysarthric audio as shown in Figure 7. This feature is crucial for increased communication for dysarthric patients, reducing the impact of dysarthria on their daily lives.

The dataset was obtained from Kaggle, containing audio files of patients with and without dysarthria as well as their corresponding text. Our approach is to fine-tune an existing speech-to-text model on audio files of patients with dysarthria,

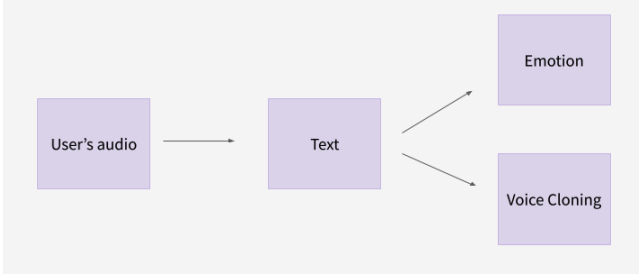


Fig. 6. Overall Speech to text, voice cloning, and emotion pipeline

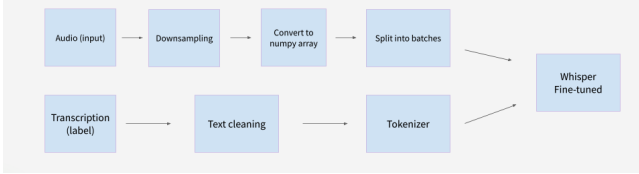


Fig. 7. Preprocessing pipeline for fine-tuning the Whisper model. The audio input undergoes downsampling, conversion to NumPy arrays, and batching. Corresponding transcriptions are cleaned and tokenized. Both processed inputs are then used to fine-tune the Whisper model for improved speech recognition performance.

thereby increasing the ability of these existing frameworks to comprehend slurred and slowed speech.

After matching each audio file to their corresponding text, all instances of non-dysarthric patients were dropped to ensure that the model was only fine-tuned on patients with dysarthria, as the chosen models already performed well on clean speech.

Three speech-to-text models were chosen for this application: Wave2Vec, Whisper, and Whisper Tiny. Transcriptions were cleaned to remove brackets and unnecessary spaces, and audio files were converted to numpy arrays and split into batches for quicker processing. The dataset was split with test size as 0.1, and transcriptions were converted using a tokenizer. Fine-tuning on Wave2Vec and Whisper proved to be difficult due to computational constraints and a large inference time, so Whisper Tiny was used for final model fine-tuning.

2) *Emotion Classification*: After obtaining speech to text results, an important component was adding an emotion classifier as shown in Figure 6, as sentiment is often lost in their reduced speech intelligibility and limitation in expressing nonverbal information [24]. We used the pretrained Emotion English DistilRoBERTa-base transformer [25] to classify english text into 7 sentiments: anger, disgust, fear, joy, sadness, surprise, and neutral. Audio recordings were converted to text using the speech-to-text converter and then subsequently passed into the DistilRoBERTa-base model to infer the speaker’s emotional state.

3) *Voice Cloning*: For patients with dysarthria, preserving their voice identity prior to their diseases is often crucial for better communication and understanding. Voice cloning is a process that reproduces a given dysarthric speech using input speech tokens from audio samples prior to the patient’s dysarthria as showcased in Figure 8.

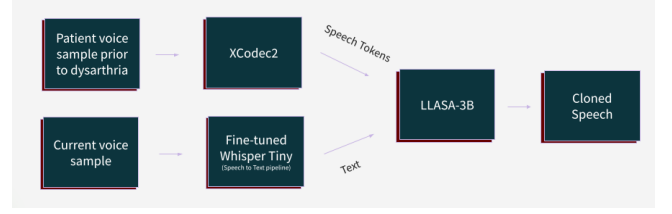


Fig. 8. Pipeline for voice cloning from dysarthric speech. The patient’s prior voice sample is encoded into speaker-specific tokens using XCodec2. Simultaneously, their current dysarthric speech is transcribed to text using a fine-tuned Whisper Tiny model. These speech tokens and text are jointly passed to the LLASA-3B model, which generates speech in the patient’s original voice.

Our voice-cloning text-to-speech (TTS) pipeline loads a user-provided sample audio (saying an arbitrary sentence) and resembles it to 16 hertz mono for input into an XCodec2 model. All reference code was taken from a pre-existing SOTA Text-to-speech and Zero Shot Voice cloning model [26]. This speech codec model encodes a speaker’s vocal characteristics in a sequence of discrete tokens. The pipeline packages the speech tokens with the output from the speech-to-text pipeline, and feeds it into a LLASA-3B model which is fine-tuned to generate speech token sequences based on the text and speaker voice tokens. The generated speech tokens are then passed back onto the XCodec2 decoded to synthesize audio.

III. RESULTS

A. Dysarthria Detection

To detect the presence of dysarthria, we trained a binary classifier on MFCC features extracted from the TORGO English dataset. The model achieved a high accuracy of 97.5%, and the training curve in Figure 9a shows stable convergence with minimal overfitting, supported by consistent validation loss.

To evaluate cross-lingual generalization, we fine-tuned the English-trained model on German and Russian datasets. The model maintained high accuracy on both these languages as shown in Table I. The confusion matrix in Figure 9b should that 98 out of 100 dysarthric and 98 out of 100 non-dysarthric samples were correctly identified in the Russian dataset, with only minimal misclassifications.

Language	Accuracy (%)
English	97.5
German	96.8
Russian	99.7

TABLE I
ACCURACY OF DYSARTHRIA DETECTION ACROSS DIFFERENT LANGUAGES.

B. Severity Classification

Our severity classifier was trained on spectrogram images of patients ranging across different severities of dysarthria. The model received a testing accuracy of 97.64% as shown in Table II. The loss curves in Figure 10 show convergence after

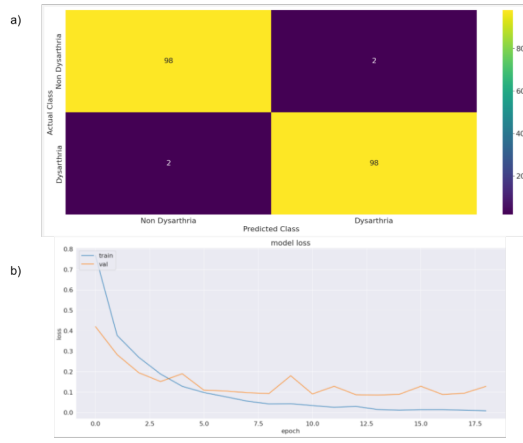


Fig. 9. (a) Training and validation loss curves for the model trained on the English (TORGO) dataset using MFCC features. The model shows stable convergence after 8 epochs. (b) Confusion matrix for the fine-tuned model on the Russian dataset, demonstrating high classification performance with 98% accuracy in both dysarthric and non-dysarthric speech classes.

8 epochs. The confusion matrix in Figure 11 shows very few misclassified instances.

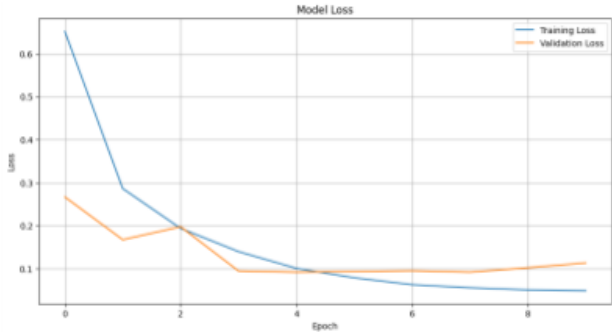


Fig. 10. Training and validation loss over epochs.

Metric	Accuracy (%)
Training Accuracy	98.18
Validation Accuracy	97.38
Testing Accuracy	97.64

TABLE II
MODEL ACCURACY ACROSS TRAINING, VALIDATION, AND TESTING DATASETS.

Another key feature of our model is interpretability. Grad-CAM heatmaps shown in Figure 12 were produced by taking the gradients of the target class score with respect to the feature maps from a convolutional layer. Regions of importance are highlighted by the image are shown in yellow and green.

C. Speech to Speech Pipelines

1) *Dysarthria Clean Speech Generation (Russian)*: Figure 13 above displays the input dysarthric spectrograms, the U-Net model's predicted outputs, and the corresponding ground truth normal spectrograms. These visualizations confirm that

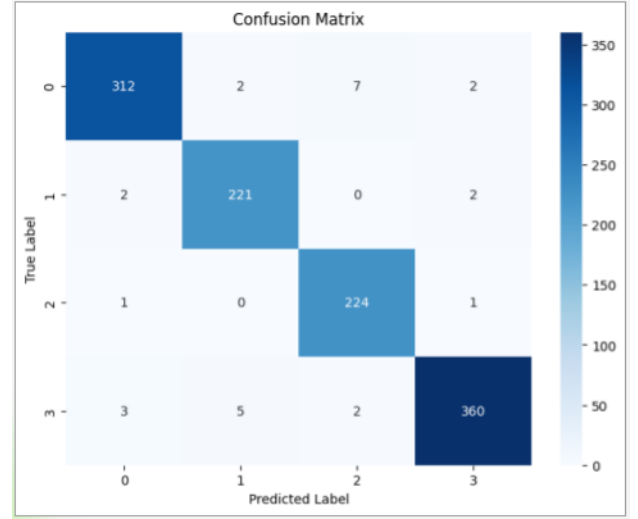


Fig. 11. Confusion matrix for severity classification.

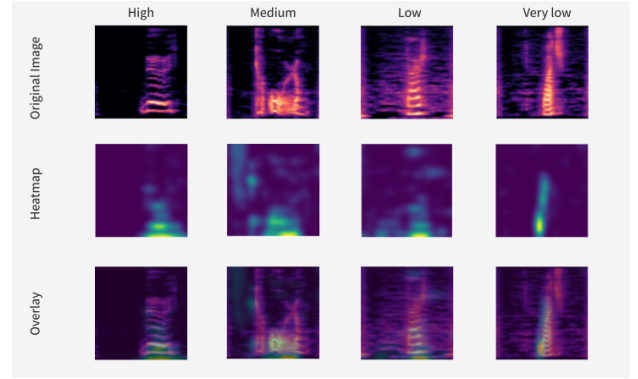


Fig. 12. Saliency Heatmap Showcasing Regions of Importance in Severity Classification

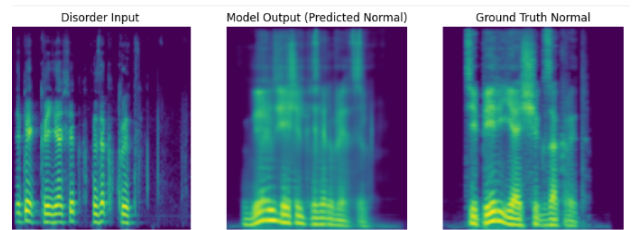


Fig. 13. Speech-to-speech transformation using Russian data. The left spectrogram shows dysarthric input speech, the middle displays the model's predicted normal output, and the right shows the ground truth normal speech. The model output recovers key time-frequency structures associated with clarity and articulation.

the model successfully learns to denoise and restructure distorted speech patterns. While the outputs preserve the broad frequency distribution of the clean speech, they exhibit slight smoothing and blurring, likely due to the limited phase reconstruction during waveform conversion.

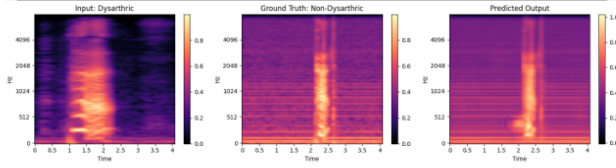


Fig. 14. Speech-to-speech transformation using English data. The left spectrogram shows input dysarthric speech, the center depicts the ground truth normal speech, and the right displays the model’s predicted output. Despite limited training data, the model captures key spectral features, yielding a clearer and more intelligible output.

2) *Dysarthria Clean Speech Generation (Extended to English)*: Figure 14 shows the results of the pretrained Russian U-Net model to generate normal speech from dysarthric speech fine tuned on the small English dataset.

In Phase 2, the Russian-trained U-Net model was fine-tuned on a much smaller, carefully filtered English dataset. The figure showcases the input English dysarthric spectrograms, model predictions, and their matched ground truth normal counterparts.

3) *Comparison of Speech to Speech Results*: The model achieved a relatively low L1 training loss (0.03) and validation loss (0.06) shown in Table III. The fine-tuned model for the English dataset achieved a low train and test loss of 0.02 loss of 0.03 respectively.

Model	U-Net (Russian)	Pretrained Russian U-Net finetuned on English
Training Loss	0.03	0.02
Test Loss	0.06	0.03

TABLE III

COMPARISON OF TRAINING AND TEST L1 LOSSES ACROSS THREE MODEL SETUPS: (1) U-NET TRAINED ON RUSSIAN PAIRED DYSARTHIC–CLEAN SPECTROGRAMS, (2) RUSSIAN-TRAINED U-NET FINE-TUNED ON ENGLISH DATA. TRANSFER LEARNING VIA RUSSIAN PRETRAINING LEADS TO GOOD PERFORMANCE ON LIMITED ENGLISH DATA.

D. Speech to Text

Our Speech-to-Text model received its best accuracy after 3 epochs. Table IV shows a training loss (word error rate) of 0.1367 towards the 3rd epoch, indicating an accuracy of 87.33%. Figure 15 indicates decreasing loss in only three epochs, a result of Whisper Tiny’s light framework designed for slurred speech. The output transcription can then be use for patient voice cloning, recreating the speaker’s original voice identity in what they say after being diagnosed with dysarthria.

E. Emotion

Patients with dysarthria often have altered sentiment in their voice due to speech impairment. Model confidences shown in table V indicate moderate levels of confidence across

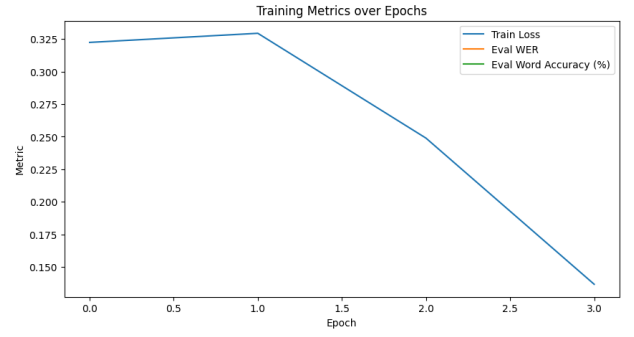


Fig. 15. Speech to Text Word Error Rate over Epochs

Epoch	Word Error Rate (WER)
0	0.3224
1	0.3294
2	0.2489
3	0.1367

TABLE IV

WORD ERROR RATE (WER) ACROSS EPOCHS DURING TRAINING.

all emotions when analyzing speech transcriptions. Table VI showcases sample sentences along with their corresponding emotion.

Emotion	Model Confidence
Anger	0.619131
Disgust	0.675264
Fear	0.625634
Joy	0.789678
Neutral	0.724674
Sadness	0.645108
Surprise	0.575870

TABLE V

CONFIDENCE SCORES FOR EACH EMOTION PREDICTED BY THE DISTILROBERTA MODEL

Sentence	Predicted Emotion
The snow blew into large drifts.	Anger
Don't ask me to carry an oily rag like that.	Disgust
Before Thursday's exam, review every formula.	Fear
Bright sunshine shimmers on the ocean.	Joy
The store serves meals every day.	Neutral
The family requests that flowers be omitted.	Sadness
Yet he still thinks as swiftly as ever.	Surprise

TABLE VI

SAMPLE SENTENCES CORRESPONDING TO EACH EMOTION

IV. DISCUSSION

Overall, we were able to achieve high classification accuracy across English, Russian, and German demonstrating that our model is able to capture the acoustic patterns of Dysarthric speech while also generalizing to cross linguistic patterns. This cross-linguistic robustness is promising for improved dysarthria classification in other languages without as much available data. Using a CNN for spectrogram-based severity detection also yielded promising and interpretable results, which enables early detection of dysarthria that a human

examiner may not be able to pick up. Saliency heatmaps show that across all classes, the model is focusing on the lower harmonics, centered around the timepoints and frequencies that the signal lies in. This confirms the reason behind model prediction, as our model is able to look at a signal with interpretable results.

The results from our speech to speech model suggests that U-Net based spectrogram translation is promising for translating dysarthric speech to normal speech. By directly learning mappings from disordered to normalized speech spectrograms, the model is able to recover key time-frequency structures associated with clarity. While most existing systems rely on large matched datasets, our approach shows that using a pretrained architecture trained on Russian speech can be applied to low-resource languages. This highlights the potential of transfer learning for low resource languages, where collecting large paired datasets may not be feasible.

ASR using transfer learning with Whisper Tiny effectively used the pretrained model and adjusted well to the limited dysarthria data using freezing and data augmentation. Whisper Tiny supports 99 languages and further research is needed for fine-tuning the transfer learning model to generalize to multiple languages, improving accessibility [27]. The main benefits of Whisper Tiny is that it is very robust, trained on 680,000 hours of multilingual data while also providing faster inferences than the two other transfer learning models tested (Wav2Vec and Whisper Tiny) [28].

Results demonstrate the effectiveness of our voice-cloning pipeline in reconstructing a patient’s original voice identity. While there is a significant degradation in voice identity between original voice and dysarthric voice, further improvements can be made by incorporating phonetic patterns such as how a speaker stresses syllables or transitions between vowels to improve voice intelligibility.

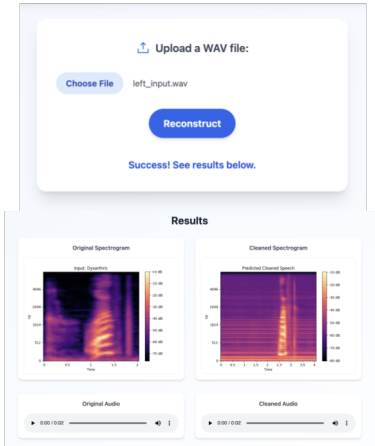


Fig. 16. Dysarthria Detection App built using HTML

A. Limitations

While our framework demonstrates promising results across multiple tasks and languages, it has some limitations. First,

our English speech-to-speech model was fine-tuned on a small paired dataset, which may restrict generalization across broader accents or sentence structures. Second, the emotion classifier was trained solely on clean transcriptions and may not fully capture emotion from more spontaneous or emotionally complex utterances. Finally, while our cross-lingual transfer approach worked well from Russian to English, its effectiveness across more structurally distant language pairs remains untested. Future work includes expanding our dataset to include more diverse speakers and dialects, improving robustness to spontaneous speech, and testing the framework on additional low-resource languages. We also aim to refine the voice cloning pipeline to better preserve speaker identity over longer and more variable inputs.

B. Mobile Launch

To maximize accessibility we developed a web app using Flask and Javascript for backend and HTML and Tailwind CSS frontend. Images of the app interface are shown in Figure 16. The Web app allows people to get timely, at-home dysarthria diagnosis results and easily use communication-aiding tools. In the future we hope to also develop a mobile app to improve accessibility.

V. CONCLUSIONS AND FUTURE WORK

In this work, we present a multilingual framework for addressing the many dimensions of dysarthria, including detection, severity classification, speech-to-text transcription, clean speech generation, emotion classification, and voice cloning. Our models show high performance across English, Russian, and German datasets, demonstrating the potential for use in real-world multilingual settings. To expand the reach of our framework, our next goal is to incorporate more low-resource languages where dysarthria diagnosis tools are especially scarce. We also aim to further reduce the Word Error Rate (WER) in our speech-to-text module by increasing dataset size, fine-tuning on more speech, and exploring multimodal data (e.g. combining acoustic features with visual inputs such as lip movements) to improve transcription accuracy. Our work is the foundation for a globally inclusive system for speech-based assistive technologies to bridge linguistic gaps and support communication and care for all patients with dysarthria.

ACKNOWLEDGMENT

We are grateful to the Medlytics instructors and teaching assistants and MIT Beaver Works Summer Institute program directors for their guidance and encouragement throughout this project.

REFERENCES

- [1] Mayo Clinic Staff, “Dysarthria: Symptoms and causes,” <https://www.mayoclinic.org/diseases-conditions/dysarthria/symptoms-causes/syc-20371994>, 2024, Accessed July 2024.
- [2] Dilip Kumar Jayaraman and Joe M Das, “Dysarthria,” in *StatPearls*. StatPearls Publishing, Treasure Island (FL), Jan. 2025.
- [3] American Speech-Language-Hearing Association, “Dysarthria in adults,” <https://www.asha.org/practiceportal/clinicaltopics/dysarthriainadults/>, 2025, Accessed August 2025.

- [4] Nick Miller, Anja Lowit, and Anja Kuschmann, "Introduction: Cross-language perspectives on motor speech disorders," in *Motor Speech Disorders: A Cross-Language Perspective*, Nick Miller and Anja Lowit, Eds., pp. 7–28. Multilingual Matters, 2014.
- [5] Yunjung Kim, Austin Thompson, and Seung Jin Lee, "Does native language matter in perceptual ratings of dysarthria?," *J. Speech Lang. Hear. Res.*, vol. 67, no. 9, pp. 2842–2855, Sept. 2024.
- [6] Esraa Hassan, Abeer Saber, Tarek Abd El-Hafeez, T. Medhat, and Mahmoud Y. Shams, "Enhanced dysarthria detection in cerebral palsy and als patients using wavenet and cnn-bilstm models: A comparative study with model interpretability," *Biomedical Signal Processing and Control*, vol. 110, pp. 108128, 2025.
- [7] Vera Wolfrum, Katharina Lehner, Stefan Heim, and Wolfram Ziegler, "Clinical assessment of communication-related speech parameters in dysarthria: The impact of perceptual adaptation," *J. Speech Lang. Hear. Res.*, vol. 66, no. 8, pp. 2622–2642, Aug. 2023.
- [8] M. Mahendran, R. Visalakshi, and S. Balaji, "Dysarthria detection using convolution neural network," *Measurement: Sensors*, vol. 30, pp. 100913, 2023.
- [9] Dong-Her Shih, Ching-Hsien Liao, Ting-Wei Wu, Xiao-Yin Xu, and Ming-Hung Shih, "Dysarthria speech detection using convolutional neural networks with gated recurrent unit," *Healthcare (Basel)*, vol. 10, no. 10, pp. 1956, Oct. 2022.
- [10] S. Sajiha et al., "Automatic dysarthria detection and severity level assessment using deep learning," *EURASIP Journal on Audio, Speech, and Music Processing*, 2024.
- [11] Seung Hee Yang and Minhwa Chung, "Improving dysarthric speech intelligibility using cycle-consistent adversarial training," 2020.
- [12] Tongzhou Wang and Yihan Lin, "CycleGAN with better cycles," 2024.
- [13] Wei-Zhong Zheng, Ji-Yan Han, Chen-Yu Chen, Yuh-Jer Chang, and Ying-Hui Lai, "Improving the efficiency of dysarthria voice conversion system based on data augmentation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 4613–4623, Nov. 2023.
- [14] Wen-Shin Hsu, Guang-Tao Lin, and Wei-Hsun Wang, "Enhancing dysarthric voice conversion with fuzzy expectation maximization in diffusion models for phoneme prediction," *Diagnostics (Basel)*, vol. 14, no. 23, Nov. 2024.
- [15] Seyed Reza Shahamiri, Vanshika Lal, and Dhvani Shah, "Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 3407–3416, 2023.
- [16] TORGO Corpus Project, "Torgo: Dysarthric and control speech corpus," <https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html>, 2009, Accessed August 2025.
- [17] Poojag718, "Dysarthria and non-dysarthria speech dataset," <https://www.kaggle.com/datasets/poojag718/dysarthria-and-nondysarthria-speech-dataset>, 2022, Accessed August 2025; originally published December 2022.
- [18] IAMHungUndji, "Dysarthria detection dataset," <https://www.kaggle.com/datasets/iamhungundji/dysarthria-detection>, 2023, Originally derived from the TORGO corpus; accessed August 2025.
- [19] Vinotha, "Ua-speech all severity dataset," <https://huggingface.co/datasets/Vinotha/uaspeechall>, 2024, Audio and text dataset of UA-Speech with severity labels.
- [20] mhantor, "Russian voice dataset (hyperkinetic dysarthria speech)," <https://www.kaggle.com/datasets/mhantor/russian-voice-dataset>, 2023, Accessed August 2025.
- [21] B. Czarnetzki, "Dysarthric german speech dataset," https://huggingface.co/datasets/B-Czarnetzki/dysarthric_german, 2023.
- [22] E. P. Balogh, B. T. Miller, and J. R. Ball, Eds., *Improving Diagnosis in Health Care*, National Academies Press, Washington, DC, 2015, Board on Health Care Services; Committee on Diagnostic Error in Health Care.
- [23] Allyson D Page and Kathryn M Yorkston, "Communicative participation in dysarthria: Perspectives for management," *Brain Sci.*, vol. 12, no. 4, pp. 420, Mar. 2022.
- [24] Lubna Alhinti, Stuart Cunningham, and Heidi Christensen, "Recognising emotions in dysarthric speech using typical speech data," in *Interspeech 2020*, 2020, pp. 4821–4825.
- [25] Jochen Hartmann, "Emotion english distilroberta-base," <https://huggingface.co/jhartmann/emotionenglishdistilrobertabase/>, 2022.
- [26] Srinivas Billa, "The SOTA text-to-speech and zero shot voice cloning model that no one knows about," Hugging Face Blog, 2025, LLaSA TTS model blog post.
- [27] "What is OpenAI whisper?," <https://www.gladia.io/blog/what-is-openai-whisper>, Accessed: 2025-8-6.
- [28] Edit a Transcription, "The whisper models," <https://whisper.net/reference/models/>, Accessed: 2025-8-6.