Automatic Speech Recognition for Dysarthric Speech

Team 9

Eshan, Kavish, Samyaka

Dysarthria is a motor speech disorder caused by neurological damage to the part of the brain that controls the physical production of speech.

50 - 90% of people with CP are affected by dysarthria

> 80% of people with ALS are affected by dysarthria

Enderby, P. (2013). Disorders of communication. Handbook of Clinical Neurology, 273–281. https://doi.org/10.1016/b978-0-444-52901-5.00022-8

Problem Statement

How can we help people with dysarthria better communicate using Automatic Speech Recognition (ASR) models?

Significance

- Dysarthric speakers can formulate syntactically correct sentences, but difficulty in pronunciation renders it un-intelligible.
- People with dysarthria are often affected by other physical handicaps, limiting their ability to interact with computers, keyboards and the environment.
- Therefore, people with dysarthria benefit greatly from automatic speech recognition (ASR) models.
- Severely dysarthric subjects may have a word-error rate of 97.5% on modern systems against 15.5% for the general population

Dataset - TORGO

All data was recorded between 2008 and 2010 in Toronto, Canada. All participants provided informed consent.

01

Eight individuals with speech impediments caused by cerebral palsy or amyotrophic lateral sclerosis and age- and gender-matched control subjects.

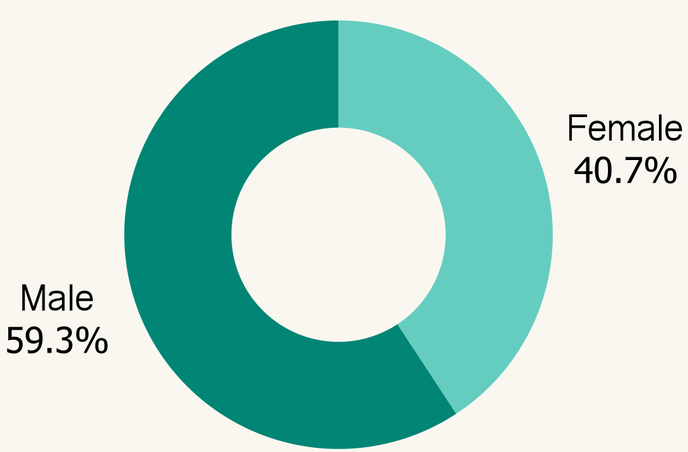
02

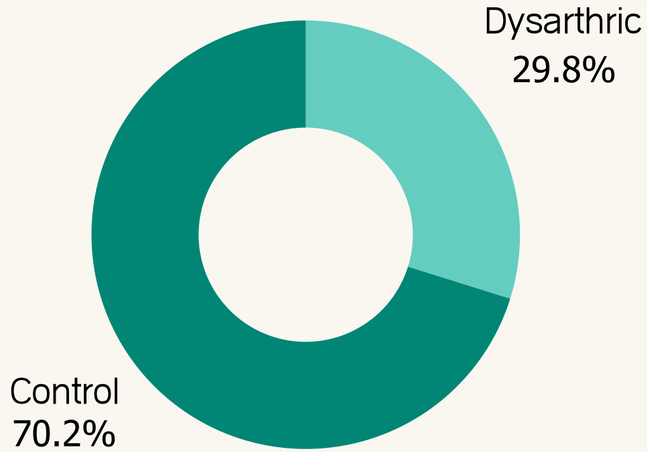
Equal number of dysarthric and control speakers useful for comparing differences, and for analyzing these relationships mathematically and functionally

03

Contains 876min or 14.6 hours of labelled data which includes:

- Non-words
- Short words
- Restricted sentences
- Unrestricted sentences





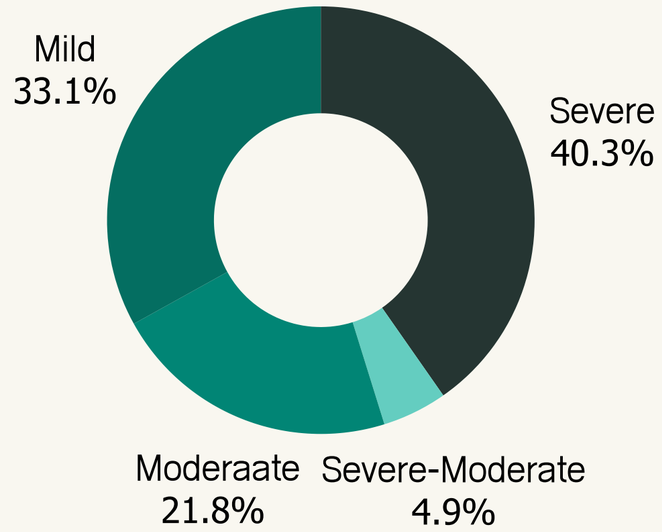


Table I
DETAILS OF TORGO DYSARTHRIC SPEECH DATABASE

Speaker	Dysarthric Speakers								Control Speakers						
	F01	M01	M02	M04	M05	F03	F04	M03	FC01	FC02	FC03	MC01	MC02	MC03	MC04
Disorder	Severe	Severe	Severe	Severe	S-M	Moderate	Mild	Mild	None	None	None	None	None	None	None
#Utterance	228	739	772	659	610	1097	675	806	296	2183	1924	2141	1112	1661	1614

*F: female speaker, M: male speaker, FC: female control speaker, MC: male control speaker; S-M represents severe-moderate category of dysarthria patients

Rudzicz, F., Namasivayam, A.K., Wolff, T. (2012) The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Language Resources and Evaluation, 46(4), pages 523--541

Dataset - LibriSpeech

01

LibriSpeech is a corpus of approximately 1000 hours of 16kHz read English speech. The data is derived from read audiobooks from the LibriVox project. (We're using about 50 hours.)

02

The dataset comprises of approximately 2484 unique speakers with 1283 of them being female speakers and 1201 male speakers.

03

Includes n-gram language models and corresponding texts excerpted from Project Gutenberg books, containing 803 million tokens and 977,000 unique words.

V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5206-5210

An Overview of Dysarthric Speech ASR Research

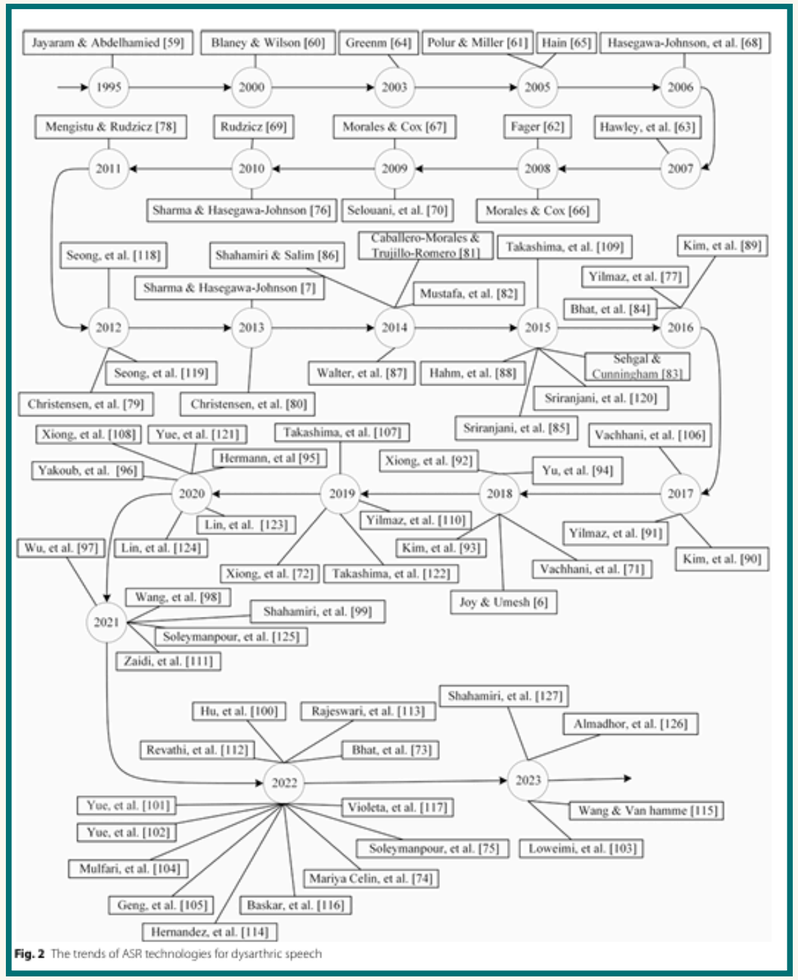


Fig. 2 The trends of ASR technologies for dysarthric speech

- Initial research explored artificial neural networks (ANN) for simple acoustic modelling and tested them on dysarthric speech.
- Researchers have also worked on various methods to go about ASR for dysarthric speech, namely speaker-dependent, speaker-adaptive and speaker-independent models, with most historical research being in the speaker-dependent field.
- As the world moved on from traditional ML methods to deep learning, general ASR models showed a huge jump in accuracy which naturally led to research on using deep learning for dysarthric speech recognition.

Qian, Z., Xiao, K., & Yu, C. (2023). A survey of technologies for automatic Dysarthric speech recognition. EURASIP Journal on Audio Speech and Music Processing, 2023(1). https://doi.org/10.1186/s13636-023-00318-2

Improving Acoustic Models in TORGO Dysarthric Speech Database
Neethu Mariam Joy, S. Umesh (2018)

Literature Review #1

Table VII
VARYING THE NUMBER OF HIDDEN LAYERS AND NEURONS PER LAYER OF DNN-HMMS TRAINED ON FMLLR FEATURES FOR F01

#Layers	#Nodes	#Tied States	%WER
España-Bonet et al. [32]			
6	1024	1800	39.57
Proposed Tuning for number of hidden layers and nodes			
3	512	600	29.32
4	512	600	29.68
4	1024	600	28.60
4	2048	600	28.96
5	1024	600	30.22
5	2048	600	31.12
Relative Reduction in WER wrt [32]			27.72

Table X
WER FOR VARIOUS DNN MODELS, ALL TRAINED ON FMLLR FEATURES

Train Data	Severe				S-M	Mod	Mild	
	F01	M01	M02	M04	M05	F03	F04	M03
Control	50.00	77.50	56.85	84.83	81.08	40.01	16.80	10.53
Dysarthric	27.16	46.08	36.93	75.92	46.97	56.26	25.47	26.41
Combined	27.34	48.89	39.62	74.89	69.78	34.43	12.53	7.89

Feature Extraction

- Using frame length of 25ms, 13-dimensional Mel frequency cepstral coefficients (MFCC) are extracted.
- First and second order MFCCs augemented to itself to generate 39-dimensional feature vectors, normalised to speaker level. Nine consecutive frames of these features are spliced together, projected to an M-dimensional vector using linear discriminant analysis (LDA) and diagonalized by maximum likelihood linear transformation (MLLT).

Methodology

- Features optimized for each dysarthric speaker's GMM-HMM model. Optimized model then used in DNN-HMM modeling.
- Used 30-dimensional FMLLR features for training and decoding DNN models.
- Supervised training uses stochastic gradient descent with a mini-batch size of 256 frames and learning rate of 0.008. The FMLLR features were stacked over a context window of 11 frames (5) and are fed as input to the DNN.
- Reduced the complexity of the DNN by reducing the number of hidden layers and nodes.

N.M. Joy, S. Umesh, Improving acoustic models in TORGO Dysarthric speech database. IEEE Trans. Neural Syst. Rehabilitation. Eng. 26(99), 637–645 (2018).

Literature Review #2

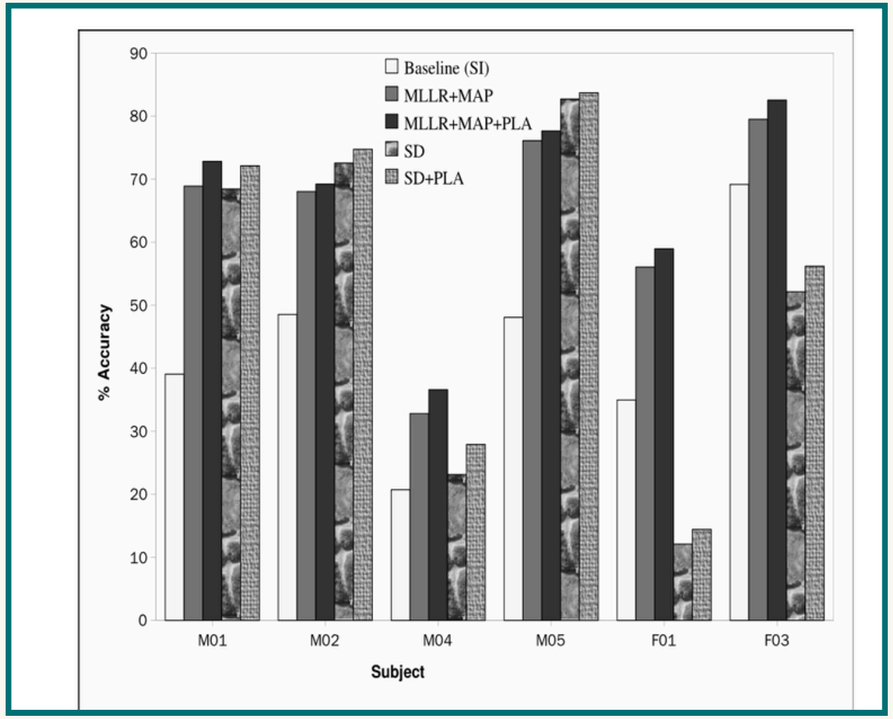
Adapting Acoustic and Lexical Models to Dysarthric Speech
Kinfe Tadesse Mengistu and Frank Rudzicz (2011)

Methodology

- HMM-GMM as baseline
- Maximum Likelihood Linear Regression (MLLR): Adapted the SI model to match individual speakers' vocal characteristics.
- Pronunciation lexicon adaptation (PLA): Since dysarthric speakers have consistent pronunciation errors, they built custom pronunciation dictionaries for each speaker.
- Different combinations of models (PLA+basline+MLLR, MLLR+PLA, etc)

Accuracy

- Control speakers had high accuracy (~85%).
- Mild-to-moderate dysarthric speakers had moderate accuracy (51-64%), but still much lower than control speakers.
- Severely dysarthric speakers had very poor accuracy (~12-30%), meaning ASR models struggle significantly with them.
- Speaker adaptation (MLLR + MAP) and pronunciation lexicons helped improve recognition, but results were still far from perfect.



Qian, Z., Xiao, K., & Yu, C. (2023). A survey of technologies for automatic Dysarthric speech recognition. EURASIP Journal on Audio Speech and Music Processing, 2023(1). https://doi.org/10.1186/s13636-023-00318-2

Literature Review #3

Enhancing English Dysarthric Speech Recognition with Age-Matched Healthy Speech: A Fine-Tuning Approach Using wav2vec 2.0
Cantao Su (2024)

Table 1: Performance of Models in Experiment 1

Model	Fine-tuning Dataset	Val WER	Evaluation WER	Evaluation CER
Model 1	Only dysarthric speech	80.28	74.96	37.71
Model 2	+ Age-matched healthy	62.24	**47.63**	**23.34**
Model 3	+ general-age healthy	71.96	60.86	30.20

Model	Intelligibility	Fine-tuning Dataset	Val WER	evaluation WER	evaluation CER
Model 4	High	+ general-age healthy	65.71	43.27	21.02
Model 5	High	+ Age-matched healthy	8.75	**35.58**	**15.91**
Model 6	Low	+ general-age healthy	81.66	**75.70**	**38.84**
Model 7	Low	+ Age-matched healthy	77.64	77.57	43.82

Table 6: Performance metrics for models in Experiment 2

Methodology

- Uses the wave2vec 2.0 model developed by facebook. Pretrained on 960h of Librispeech dataset, where each vector represents a latent feature of the speech unit.
- Fine tuned on the Torgo and Mozilla Common Voice dataset, where 3 models were used: Dysarthric speech: TORGO (≈ 200 min total).
  Age-matched healthy: TORGO controls (≈ 250 min).
  General-age healthy: Mozilla Common Voice 17.0 (≈ 250 min)

Accuracy

- For experiment 1, the best performing model (dysarthric + Age-matched healthy) achieved a WER of ~48% and a CER of ~23%.
- Experiment 2 focused on a speaker dependant fine tuned model using the best performing model from experiment 1. ~36% WER and 16% CER for highly intelligible speakers and 76% WER and 39% CER for low intelligible speakers.

Enhancing English Dysarthric Speech Recognition with Age-Matched Healthy Speech: A Fine-Tuning Approach Using wav2vec 2.0 - Studenttheses Campus Fryslan. (n.d.). Retrieved May 2, 2025, from https://campus-fryslan.studenttheses.ub.rug.nl/543/

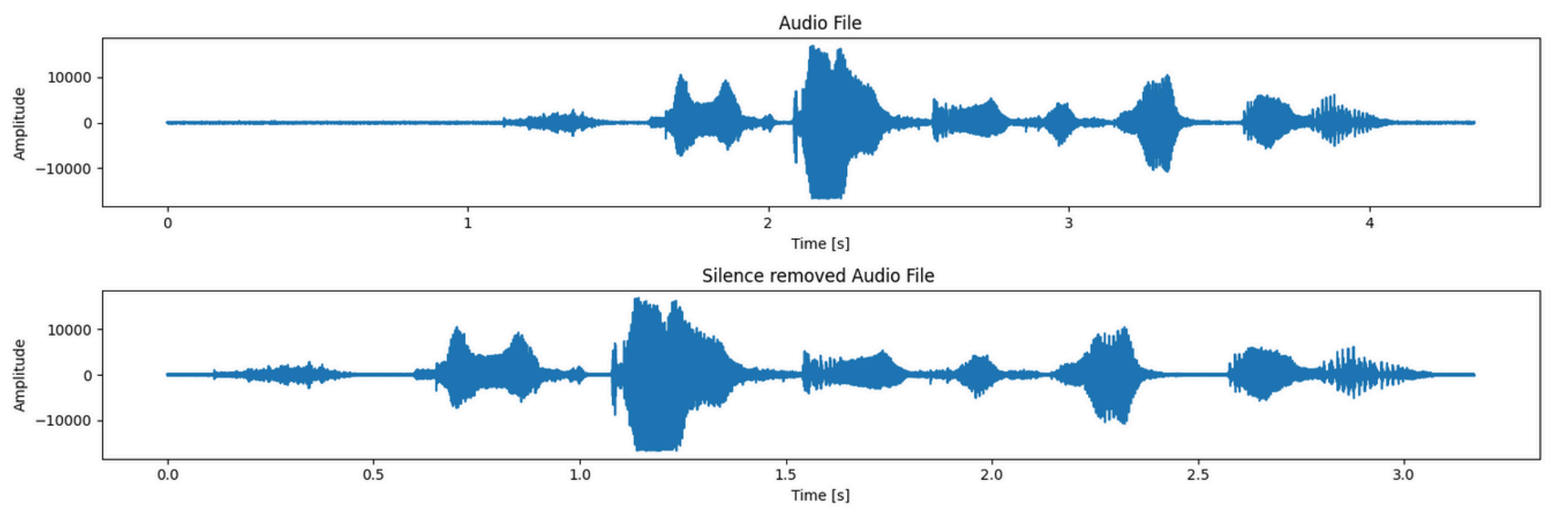
Shortcomings

- Mainly focused on speaker-dependent models and not speaker-independent models, which are crucial for real-world applications.
- Existing models are quite inaccurate with severely dysarthic speech.
- Lack of real time streaming ASR models for dysarthric speech.

Features Preprocessing

Silence Removal

Removed the silent parts from the beginning and end of the file while preserving silence between words.



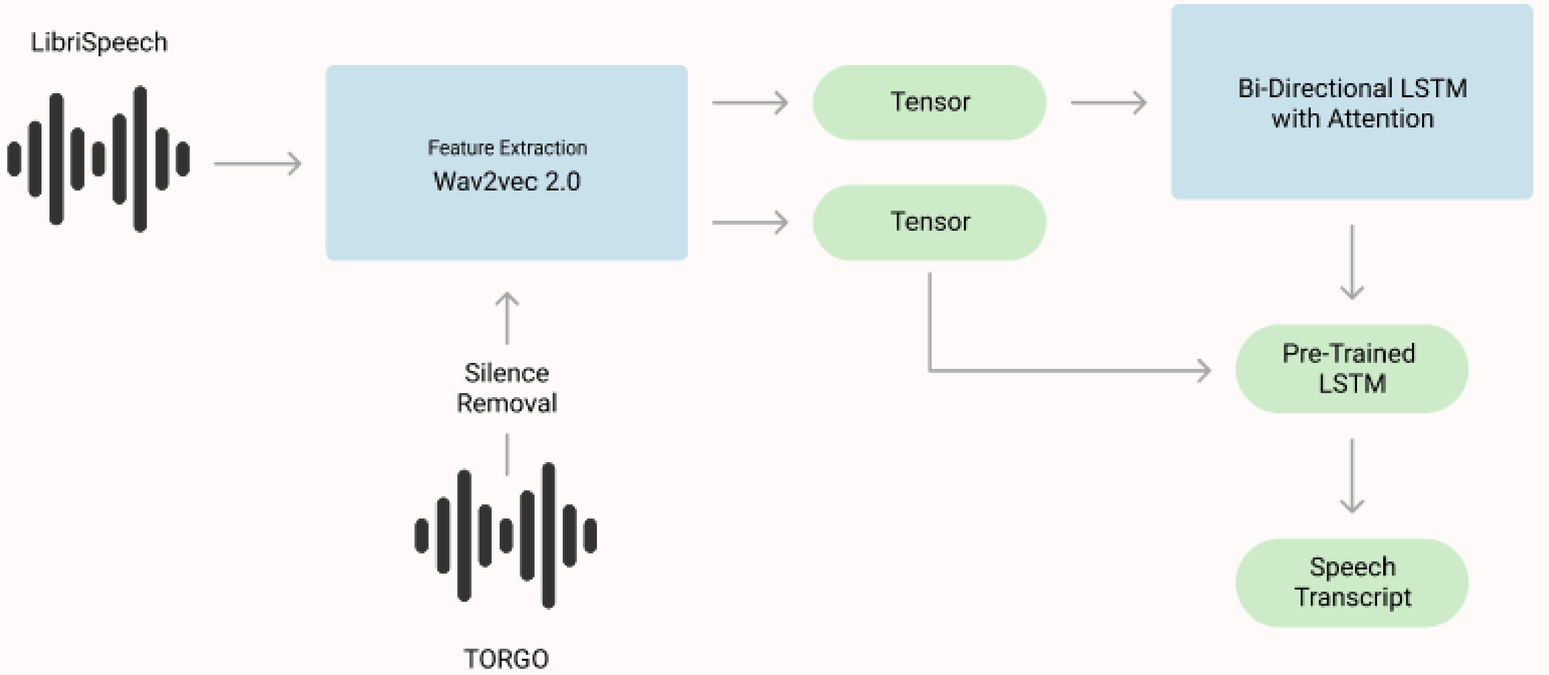
Normal vs. Silenced Audio

Methodology

All the speech from the Torgo dataset and the Librispeech dataset are first passed through the wav2vec 2.0 model for feature extraction where they are converted to 1024-d vectors.

A bi-directional LSTM model is then trained on the Librispeech data where we get an overall Word Error Rate of around 1.3%. The model is then finetuned on dysarthric speech data from the Torgo dataset where we get an overall WER of 48% and CER of 27%.

AUTOMATIC SPEECH RECOGNITION FOR DYSARTHRIC SPEECH



Our Model

- Projection Layer - 1024 to 256 dimensions
  - Normalizing
  - ReLU activation function

- Bi-directional LSTM
  - Residuals for skip connections

- Attention Layer
  - Attn. weights for each timestep using softmax activation

- Linear Layer
  - For each time step, computes logits for each character

- Trained on Connectionist Temporal Classification (CTC) Loss

Why This Model?

01

Wav2vec is able to extract high quality feature representations of the data which can model the variances in dysarthric speech.

02

A Bi-Directional LSTM is then used since we have sequential data where context matters in both directions. This is a many to many problem.

03

We are using an attention layer to help the model focus on the most important bits of the input speech.

Challenges

Limited Dataset

The Torgo dataset is limited in terms of number of speech hours for dysarthric speakers.

Unavailability

UASpeech, a larger dysarthric speech dataset wasn’t officially available.

Overfitting

Due to the limited data, our model had a really low training loss and a high validation loss even after implementing early stopping. On a larger dataset like libri, our model got a word error rate of just 1.3%

Results

Overall WER (%) and CER (%)		
Speaker	**WER**	**CER**
Overall	48.75	26.62

WER (%) and CER (%) for control speakers		
Speaker	**WER**	**CER**
FC01	53.47	30.69
FC02	42.25	17.76
FC03	46.35	26.77
MC01	39.01	17.18
MC02	46.66	24.14
MC03	26.75	10.96
MC04	32.48	15.68
WER (%) and CER (%) for dysarthric speakers		
Speaker	**WER**	**CER**
F01	93.31	60.14
F03	72.25	43.22
F04	39.38	18.4
M01	100	63.58
M02	94.91	66.62
M03	34.95	18.04
M04	92.87	62.5
M05	99.12	72.22

Potential Impact and Applications

- With a bigger dataset available, our model will potentially be able to help a large number of dysarthric speakers across the world communicate better without the need for personalized models.

- In Plaksha, anyone suffering from mild dysarthria should be able to use our model with a relatively low WER and CER

Thank You