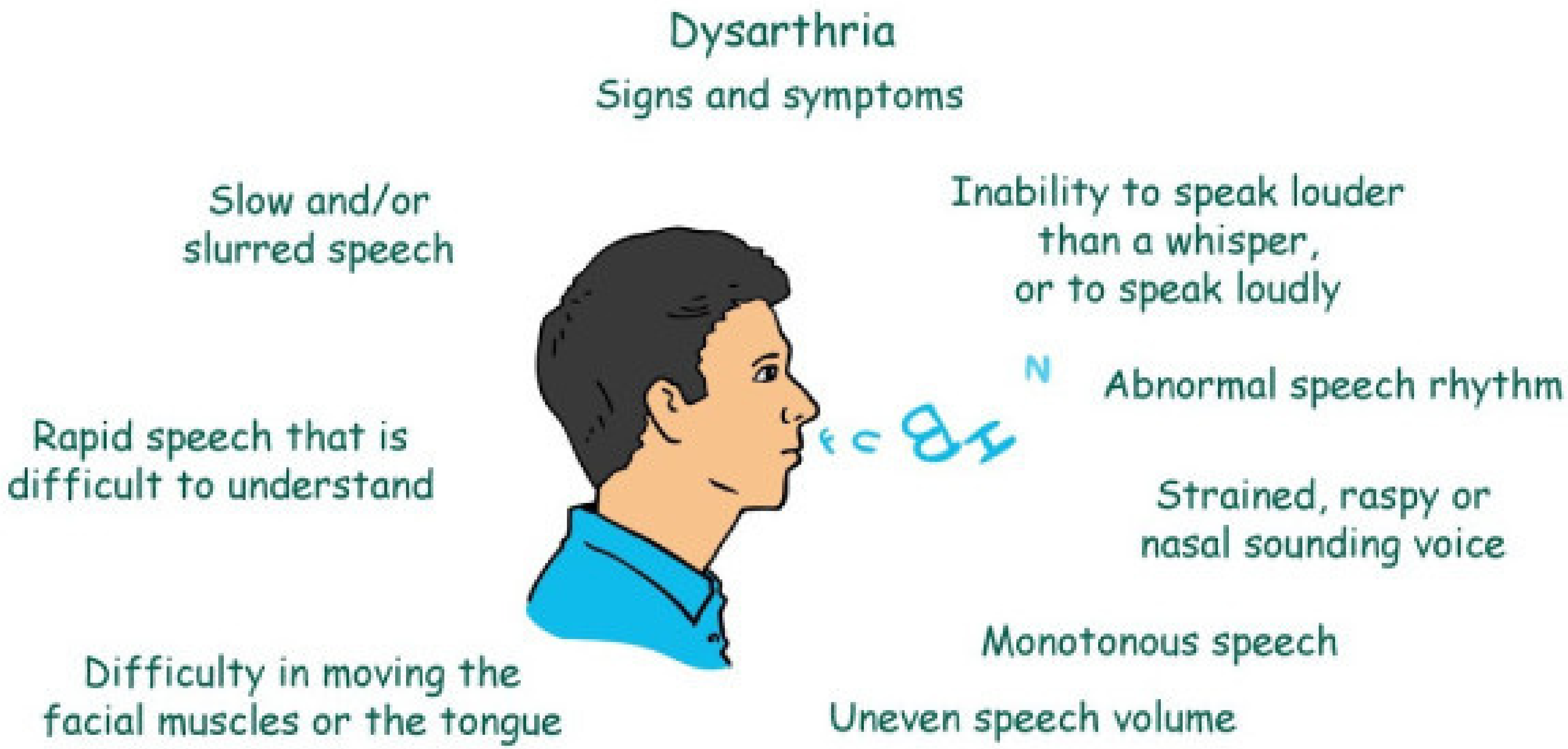
# Phonetic Analysis of Dysarthric Speech Tempo and Applications to Robust Personlised Dysarthric Speech Recognition

Feifei Xiong · Jon Barker · Heidi Christensen

IEEE ICASSP-2019

Saketh Reddy Vemula - 2022114014
Viswanath Vuppala - 2022101084

# Introduction




- Increased respiration frequency
- Inadequate Pauses
- Breathy or hoarse voice
- Reduced speech
- Deviations in pitch and volume
- Mis-articulated Sounds

# Why Phonemic Analysis?

**Phonemes in Speech Recognition:**

- Phonemes form the backbone of speech recognition systems by mapping acoustic signals to meaningful sounds (e.g., vowels and consonants).

**Challenges in Dysarthric Speech:**

- Dysarthric speakers exhibit high phonemic variability, making it difficult for systems trained on typical speech to recognize dysarthric speech accurately.
- Dysarthria is **often associated with** severe physical disabilities like **Cerebral Palsy** (conditions that affect movement and posture).
- For this group of people, **speech-enabled and hands-free interfaces** often provide a more attractive and efficient means of access in comparison to hardwired switches, keyboards and remote controls.
- Hence, ASR is one of the best options to tackle this problem.

# Approach

Typical Speech ⇄ Dysarthric Speech

- Abundant Training Data

- Very Limited Training Data

Data from UASpeech corpus → Phonetic Analysis (Forced Alignment) → Speech Tempo Ratio → Speech tempo Adjustment for ASR model trained on Typical Speech (2 stages) → Experimental Results

# Phonetic Analysis

## Speech Tempo Analysis

1. Data Selection
2. Data Preprocessing
3. Forced-alignment
4. Speech Tempo Analysis based on Phoneme Segments

CTL: Control/Typical Speech
DYS: Disorder/Dysarthric Speech

### Training Data for STA

| Sets(#Spk) | Re-segment | Block 1 & 3 | Block 2 | WER |
|---|---|---|---|---|
| CTL #13 | ✗ | 46410 (22.7 h) | 23205 (11.1 h) | 57.42 |
| | ✓ | 46403 (19.8 h) | 23205 (9.7 h) | 56.86 |
| DYS #15 | ✗ | 49204 (44.3 h) | 24731 (21.7 h) | 48.60 |
| | ✓ | 49204 (27.3 h) | 24727 (13.4 h) | 44.91 |

WER: baseline ASR performance with DYS test set

| | | |
|---|---|---|
| Vowels #16 | (V1) short vowels | AH AO AX EH IH UH |
| | (V2) medium vowels | AE |
| | (V3) long vowels | AA ER IY UW |
| | (V4) diphthongs | AW AY EY OW OY |
| Consonants #24 | (C1) glides | L R W Y |
| | (C2) unvoiced stops | K P T |
| | (C3) voiced stops | B D G |
| | (C4) nasals | M N NG |
| | (C5) unvoiced fricatives | F S SH TH |
| | (C6) voiced fricatives | DH V Z ZH |
| | (C7) unvoiced affricates | CH |
| | (C8) voiced affricates | JH |
| | (C9) aspirates | HH |

# Phonetic Analysis

## Speech Tempo Analysis

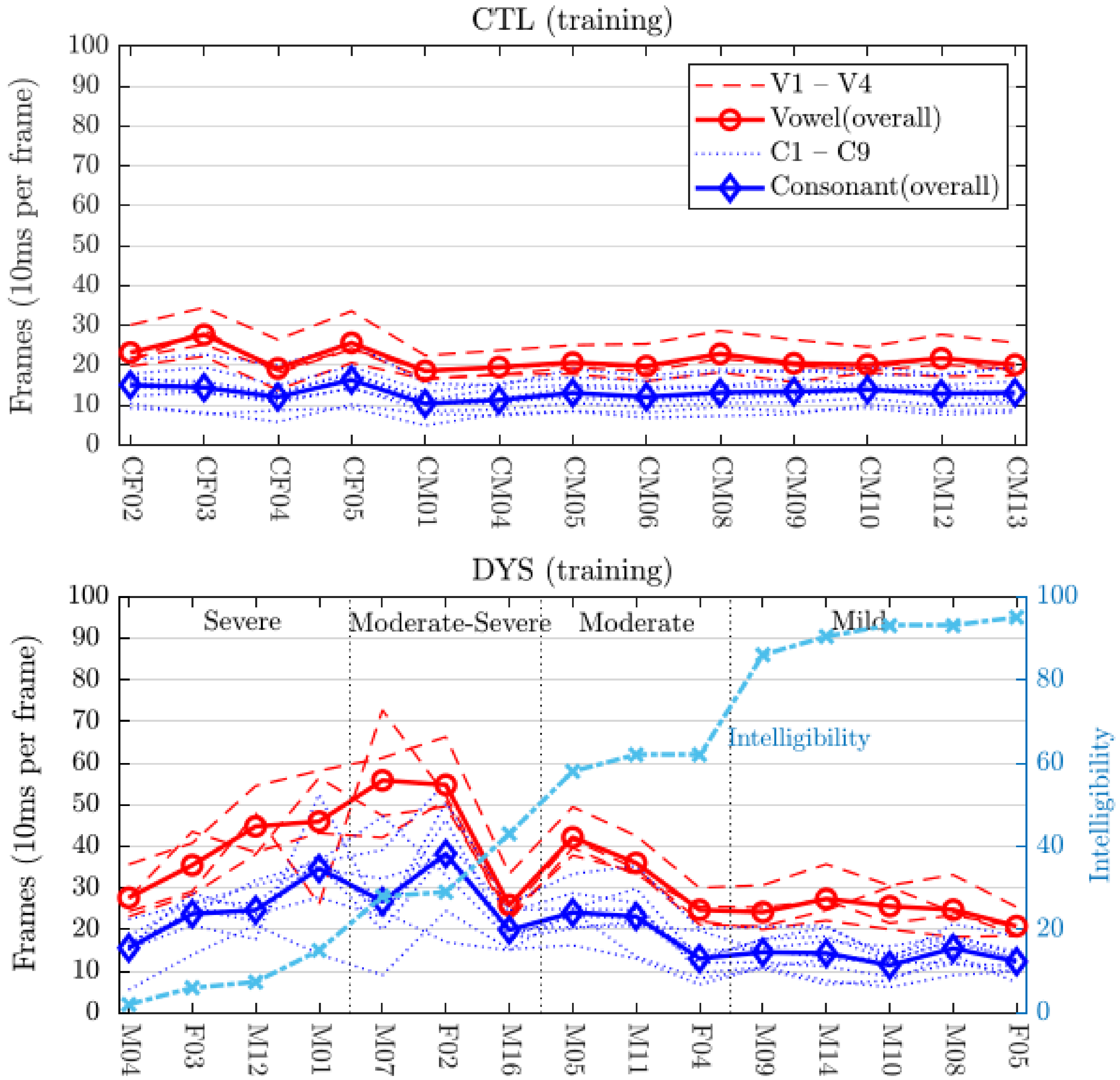
1. Data Selection
2. Data Preprocessing
3. Forced-alignment
4. Speech Tempo Analysis based on Phoneme Segments

Phoneme-based Speech Tempo Ratio:

$$\mathcal{R}_{d\leftarrow c}(p) = \frac{T_d(p)}{T_c(p)} \longrightarrow \overline{\mathcal{R}_{d\leftarrow c}}$$

CTL: Control/Typical Speech

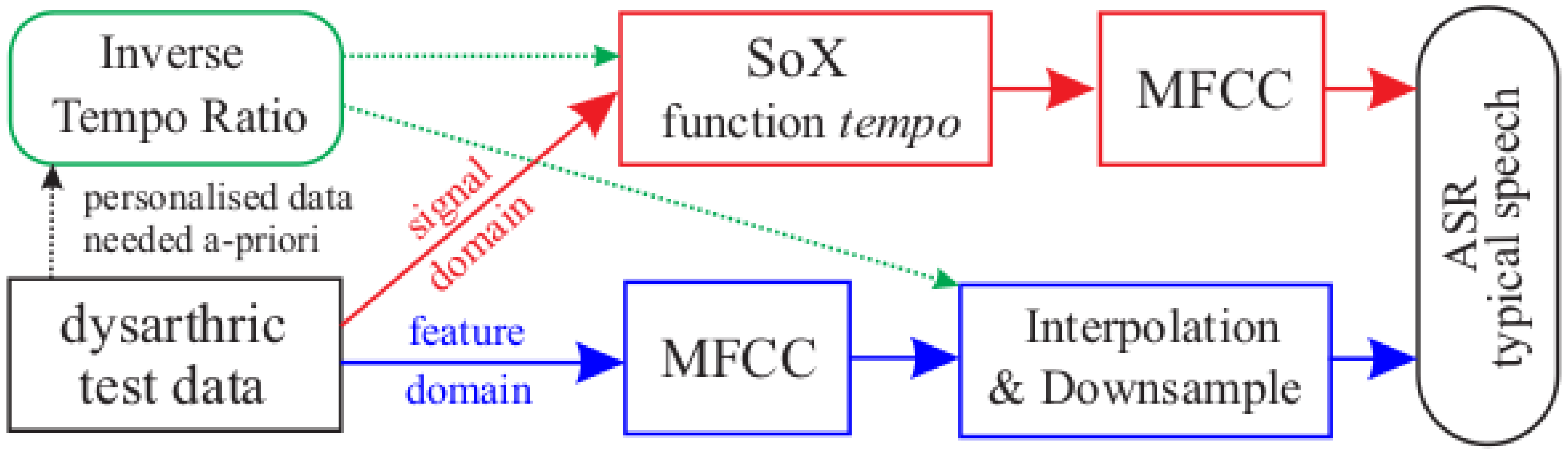
DYS: Disorder/Dysarthric Speech



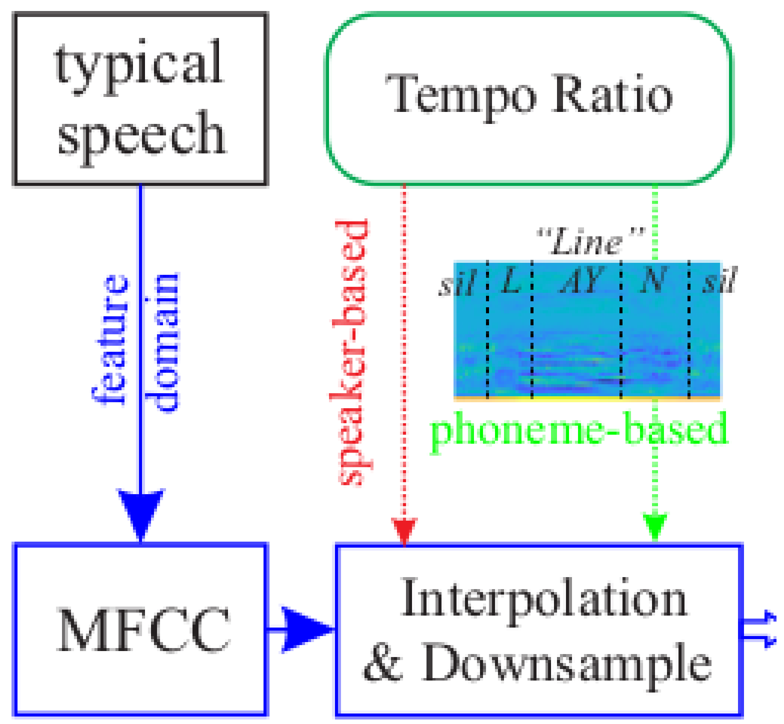
# Speech Tempo Adjustment for ASR

## Test Stage

$\overline{\mathcal{R}_{d\leftarrow c}}$



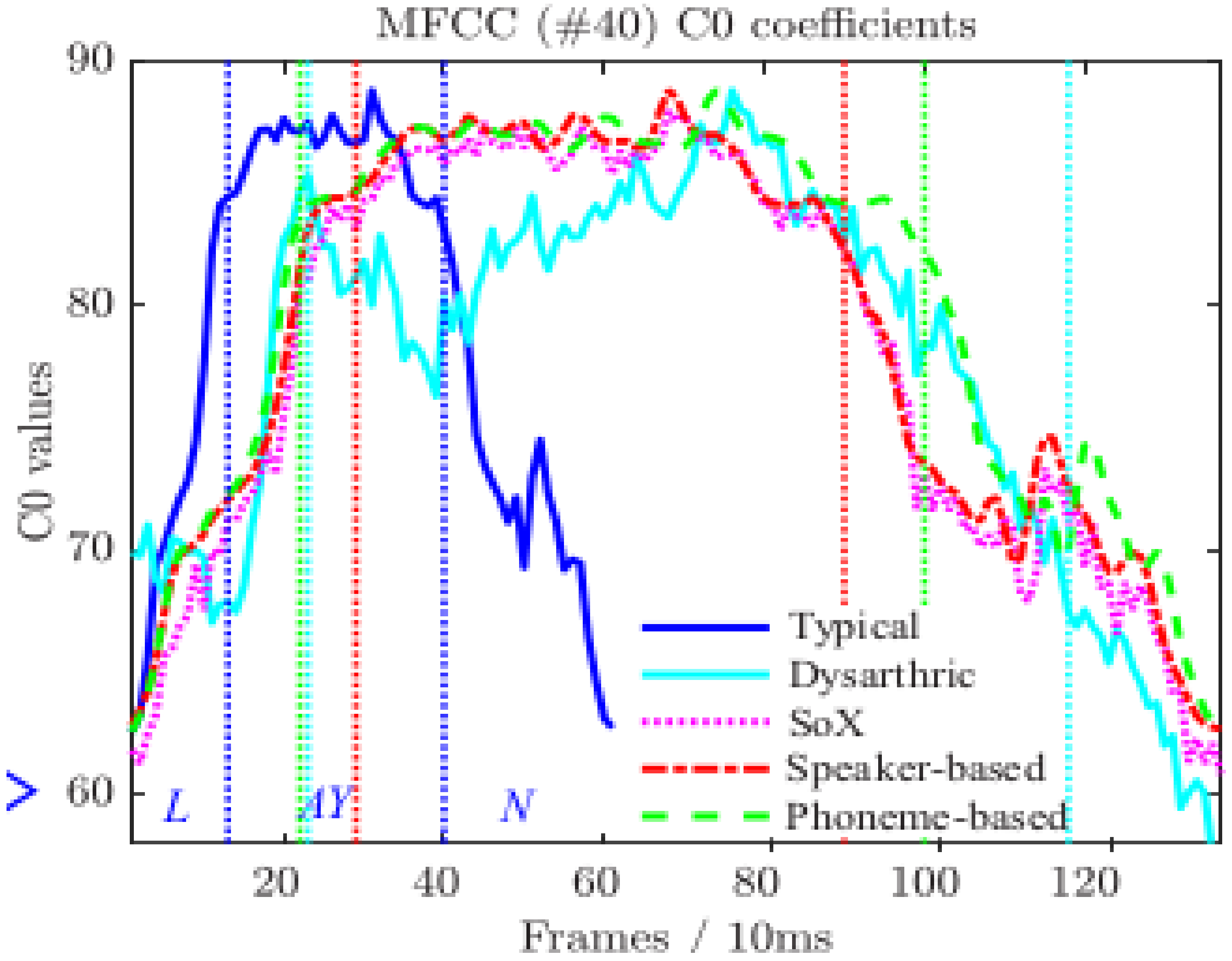

## Training Stage




$$\mathcal{R}_{d\leftarrow c}(p) = \frac{T_d(p)}{T_c(p)}$$

**Note**:

Phoneme-based tempo ratios are not possible in Test Stage due to lack of alignment knowledge in dysarthric test data



# Experimental Results

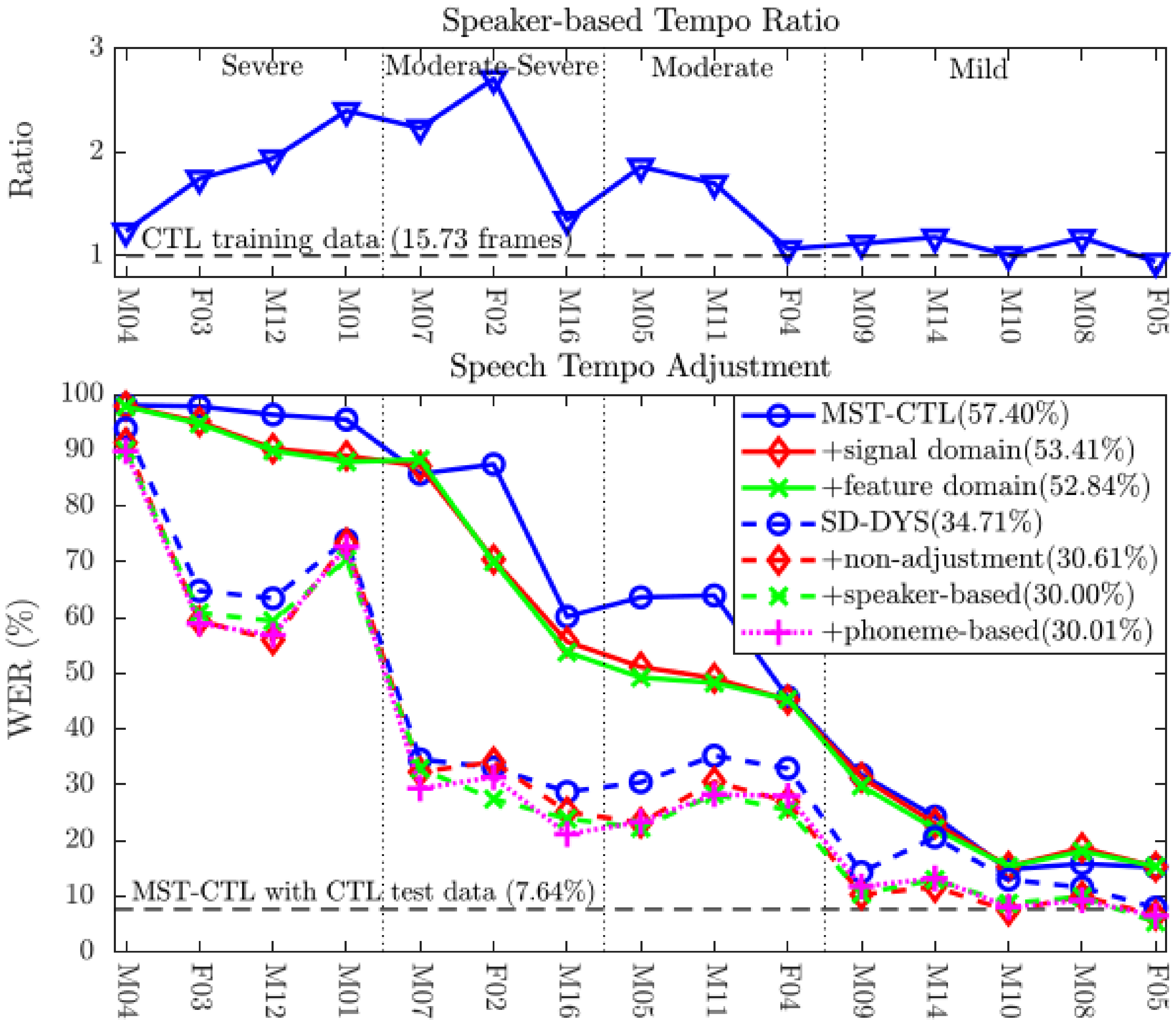
## WER vs Speech Tempo Adjustment methods

### Test Stage

1. Required to speed up withing 3 times to match
2. On Average, 4.6% absolute WER reduction can be achieved
3. Still far from SD-DYS

### Training Stage: Data Augmentation

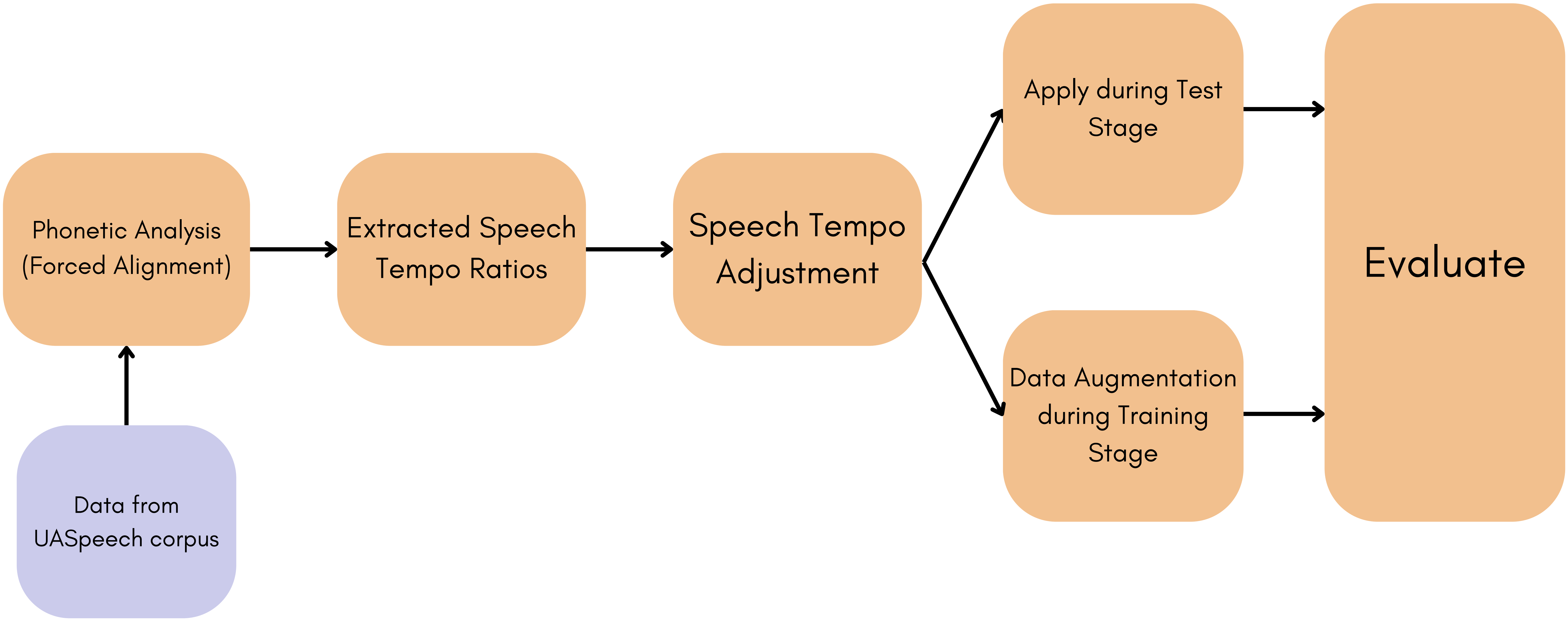
| Training | Severe | Mod.-Severe | Moderate | Mild | Overall |
|---|---|---|---|---|---|
| SD-DYS | 72.65 | 32.25 | 32.70 | 13.44 | 34.71 |
| +non-adjustment | 68.22 | 30.74 | 26.66 | **9.15** | 30.61 |
| +speaker-based | 68.76 | 28.23 | **25.13** | 9.44 | **30.00** |
| +V1−V4 | 69.33 | 29.13 | 25.47 | 9.62 | 30.50 |
| +C1−C9 | 70.69 | 29.54 | 27.41 | 9.40 | 31.16 |
| +phoneme-based | **67.83** | **27.55** | 26.41 | 9.71 | 30.01 |



# Conclusion

Presents Two-approaches for improving Dysarthric Speech Recognition

Data Augmentation strategy is more effective with **7% absolute improvement** (after including more data 3x) in comparison to baseline speaker-depended trained system.



# Thank You