CLASSIFICATION OF BISYLLABIC LEXICAL STRESS PATTERNS IN DISORDERED SPEECH USING DEEP LEARNING

Mostafa Shahin, Beena Ahmed (Texas A&M University at Qatar)

Ricardo Gutierrez-Osuna (Texas A&M University)

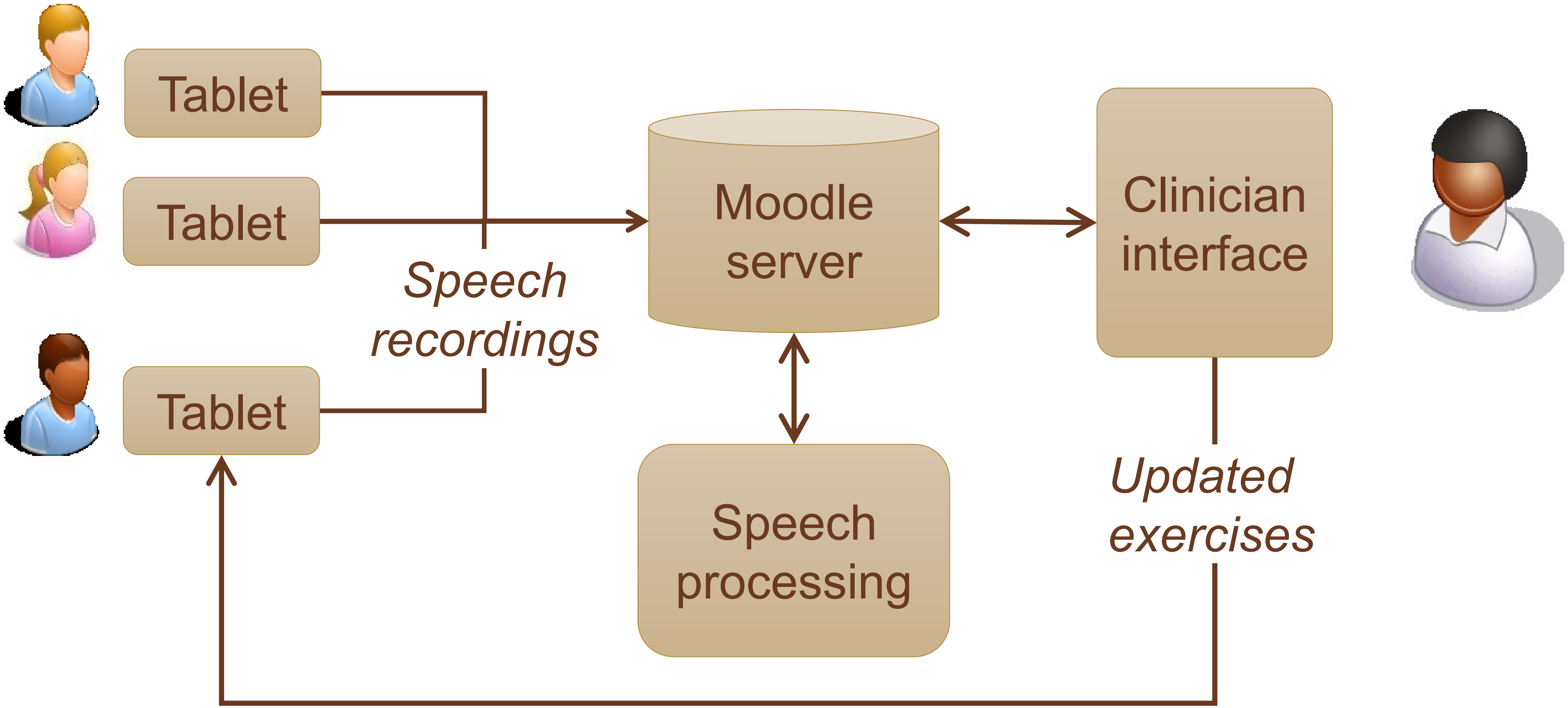
Outline

- Introduction
- Our Remote Therapy Tool
- Lexical Stress in English
- Method
- Experiments & Results
- Conclusions
- Q&A

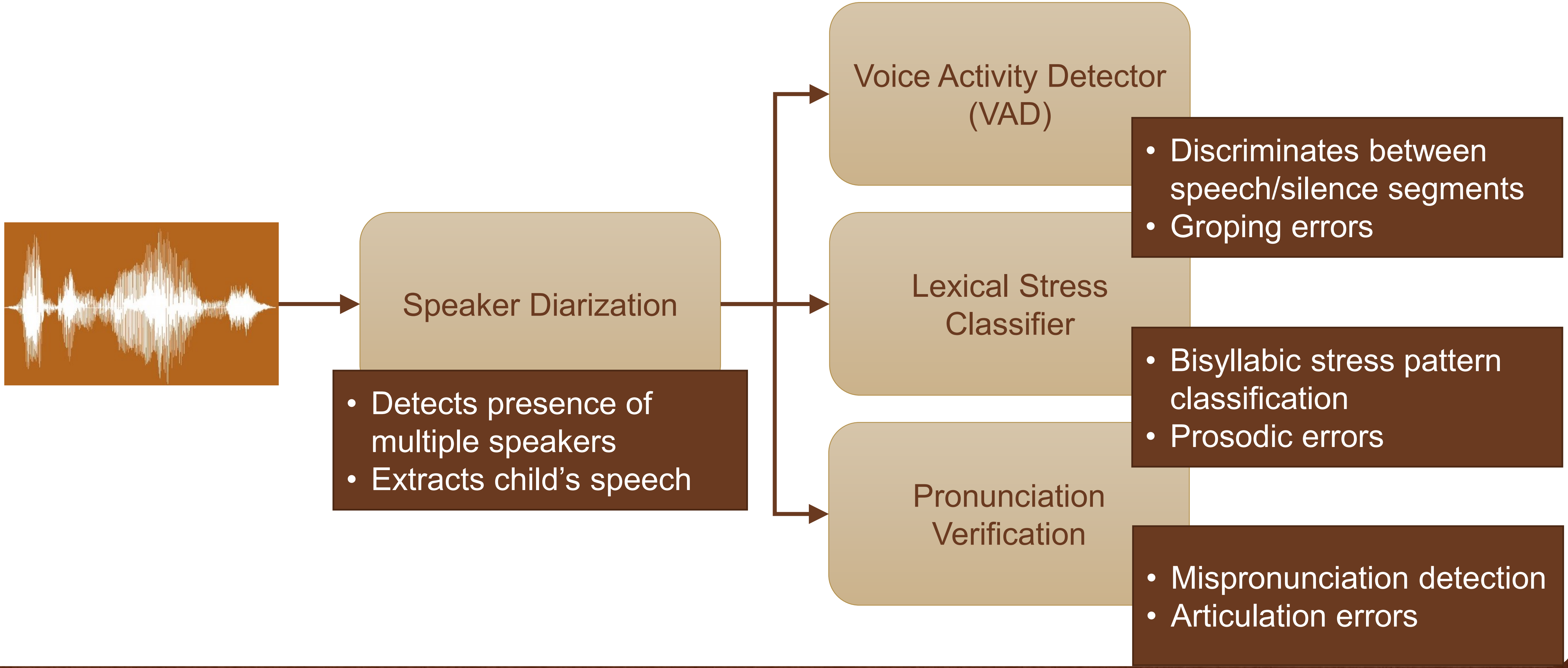
What is Childhood Apraxia of Speech (CAS)?

- Speech disorder that can lead to serious communicative disability
- Affects the ability to correctly pronounce sounds, syllables and words
- Due to neurological problems not muscular
- 3.4% - 4.3% of children in the US diagnosed with CAS.

Our Remote & Automated Therapy Tool (big picture)



Speech Processing Modules



Speaker Diarization

- Detects presence of multiple speakers
- Extracts child's speech

Voice Activity Detector (VAD)

- Discriminates between speech/silence segments
- Groping errors

Lexical Stress Classifier

- Bisyllabic stress pattern classification
- Prosodic errors

Pronunciation Verification

- Mispronunciation detection
- Articulation errors

Lexical Stress in English

- English is a stress-timed language.
- In a multi-syllabic words there is at least one stressed syllable.
- The stressed syllable can be characterized by increasing in duration, intensity and pitch.
- Pronouncing the correct stress pattern is important for the intelligibility.
- Each of two consecutive syllable has one of four possible stress patterns

SW	**TA**.ble	**FRI**.day
WS	sub.**MIT**	in.**SIDE**
SS	**CHILD.HOOD**	**FOOT.BALL**
WW	**MU**.tu.al	**OB**.vi.ous

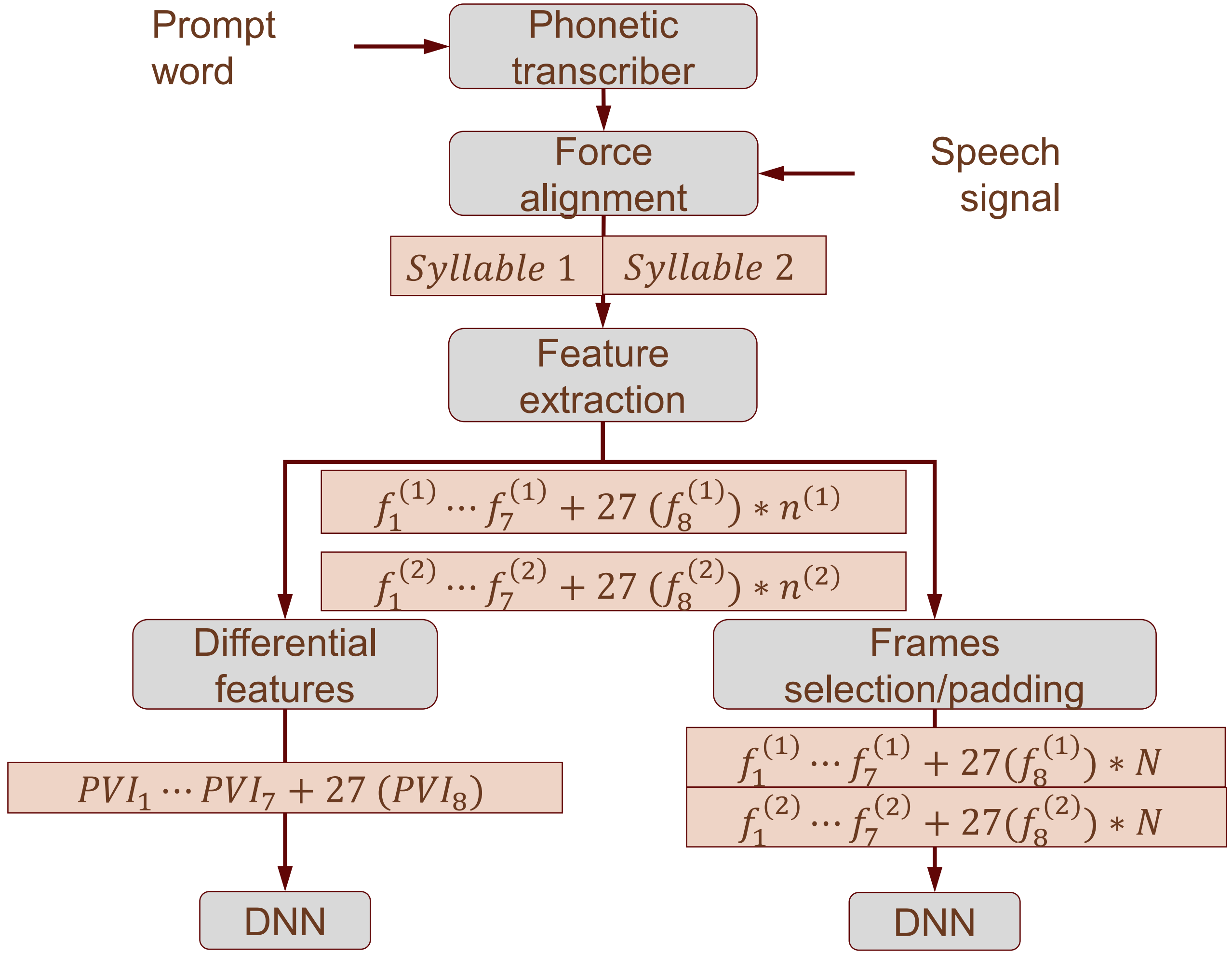
Prosodic Errors

- Children with a range of speech disorders, including childhood apraxia of speech (CAS), struggle to produce the correct lexical stress patterns.
- Incorrect production of lexical stress, i.e. prosodic errors, lead to robotic-like speech and intelligibility
- These errors are more obvious in words with unequal stress pattern, e.g. ‘banana’.
- During treatment, the therapist guides the child on how to control stress levels in pairs of adjacent syllables

Feature Extraction

Intensity	• Peak-to-peak amplitude over syllable nucleus (f_1) • Mean energy over syllable nucleus (f_2) • Maximum energy over syllable nucleus (f_3)
Pitch	• Maximum pitch over syllable nucleus (f_4) • Mean pitch over syllable nucleus (f_5)
Duration	• Nucleus duration (f_6) • Syllable duration (f_7)
Spectral	• 27 Mel-spectral energies per frame over nucleus (f_8)

System Block Diagram



Prompt word

Phonetic transcriber

Force alignment

Speech signal

$Syllable\ 1$ | $Syllable\ 2$

Feature extraction

$f_1^{(1)} \cdots f_7^{(1)} + 27\ (f_8^{(1)}) * n^{(1)}$

$f_1^{(2)} \cdots f_7^{(2)} + 27\ (f_8^{(2)}) * n^{(2)}$

Differential features

$PVI_1 \cdots PVI_7 + 27\ (PVI_8)$

DNN

Frames selection/padding

$f_1^{(1)} \cdots f_7^{(1)} + 27(f_8^{(1)}) * N$

$f_1^{(2)} \cdots f_7^{(2)} + 27(f_8^{(2)}) * N$

DNN

Differential Features

- Compute the pair-wise variability index (PVI) for each feature

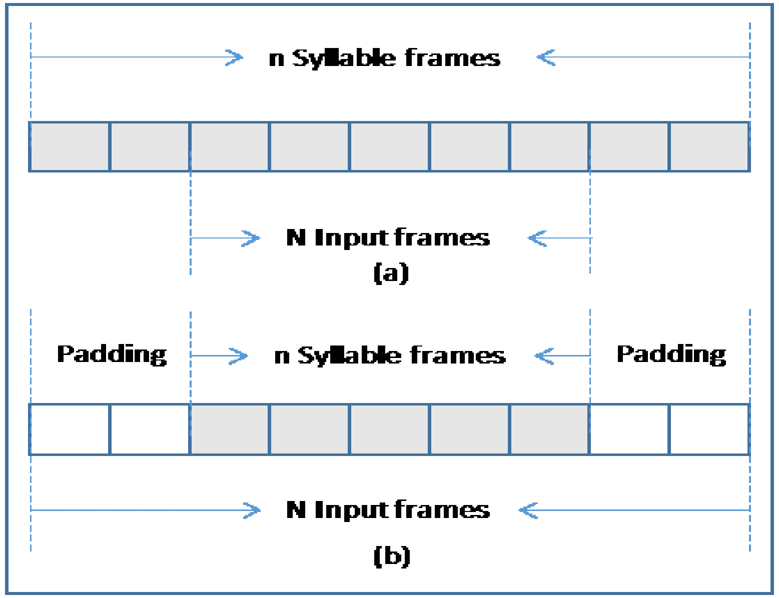
$$PVI_i = \frac{f_i^{(1)} - f_i^{(2)}}{(f_i^{(1)} + f_i^{(2)})/2}$$

$f_i^{(1)}$	The ith feature of the first syllable
$f_i^{(2)}$	The ith feature of the second syllable

- The 27 Mel-spectral energies averaged over nucleus frames to produce 27 averaged values per syllable.
- The resulted feature vector consists of 34 values representing each pair of consecutive syllables.

Raw Features

- Concatenate the extracted features of the two consecutive syllables into one wide feature vector.
- Each syllable has 7 scalar values $f_1 - f_7$ and $27 * n$ Mel-coefficients where n is the number of frames in each syllable's vowel.
- The number of frames fixed to N frames selected from middle of the vowels if n > N, or padded to N if n < N.
- The number of frames N is determined empirically.
- The size of the produced feature vector equal to: $2 * (7 + 27 * N)$



DNN Classifier

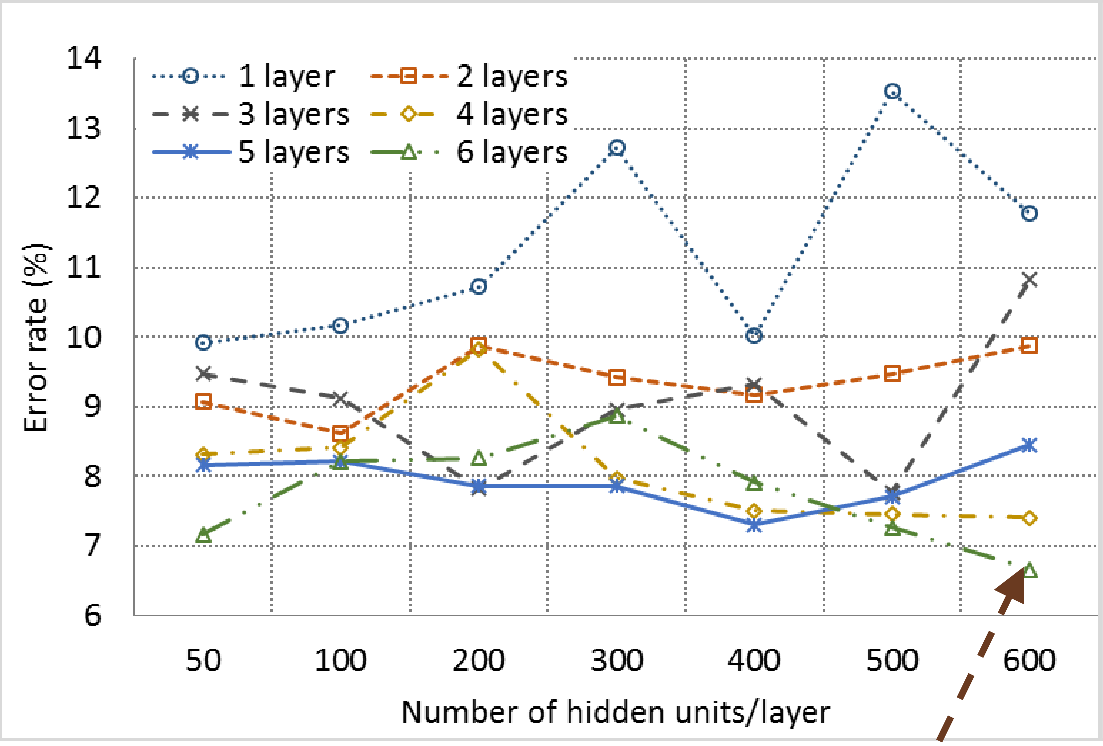
- Multi-hidden layers feedforward neural network.
- Backpropagation learning using mini-batch stochastic gradient decent method (MSGD) with adaptive learning rate.
- 4 way soft max top layer for the four possible classes (SW, WS, WW, SS).
- Tuning parameters:
  - Number of hidden layers
  - Number of hidden units per layer
  - Number of frames (N).

Speech Corpora

- Typically development speech corpus:
  - Around 500 children ranging from grade 0 to 10
  - Each child pronouncing 100 single multi-syllabic words
  - Phoneme sequence and syllable stress-level extracted automatically using CMU pronunciation dictionary
- Disordered speech corpus:
  - 10 children with CAS aged 4 - 12 years
  - Each child pronouncing 15 isolated words: 10 with a SW pattern across the first two syllables (e.g., DInosaur) and 5 with a WS pattern (e.g., toMAto)
  - The stress-level of each syllable marked manually by SLP

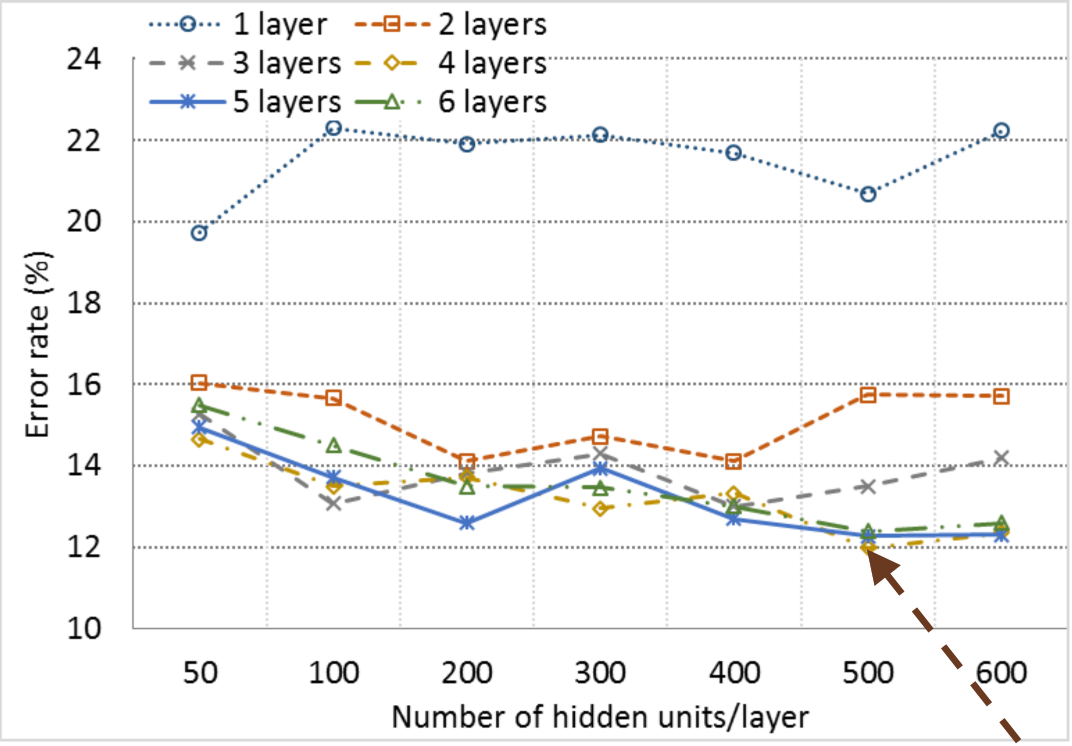
Raw feature DNN (typically development corpus)

Fixed frame size (N) of 25 frames



SW/WS

Best ER 6.8% @ 6 layers/600 units

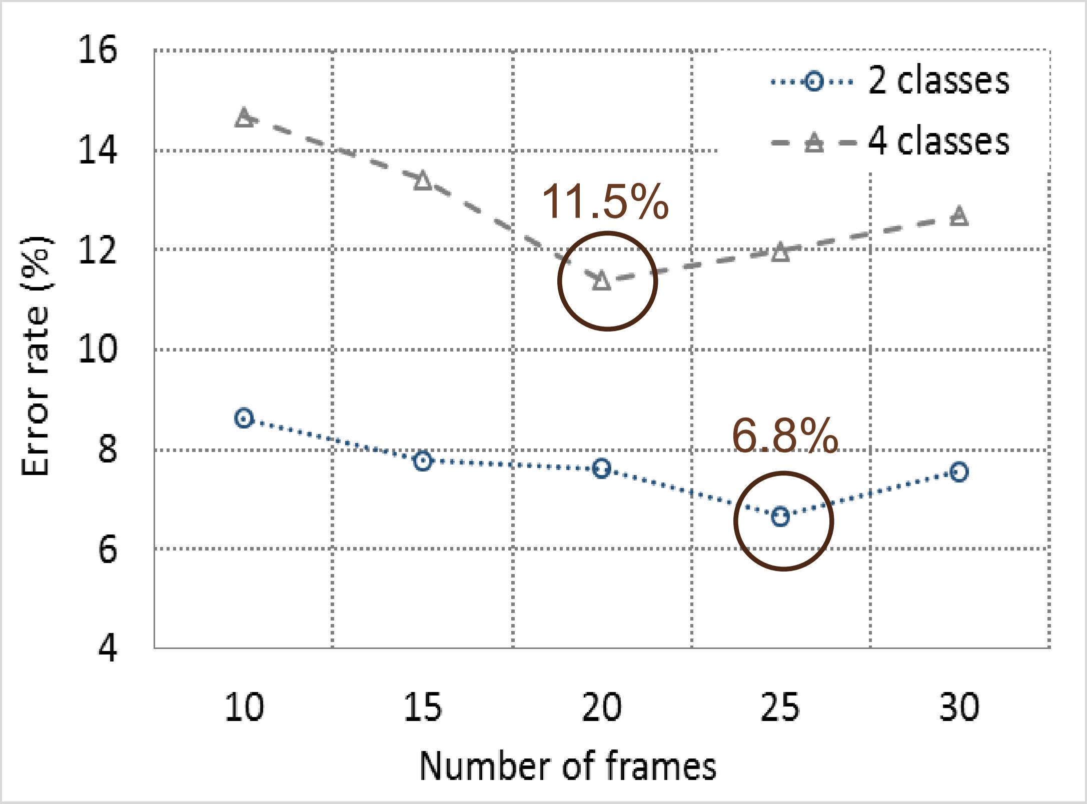


SW/WS/SS/WW

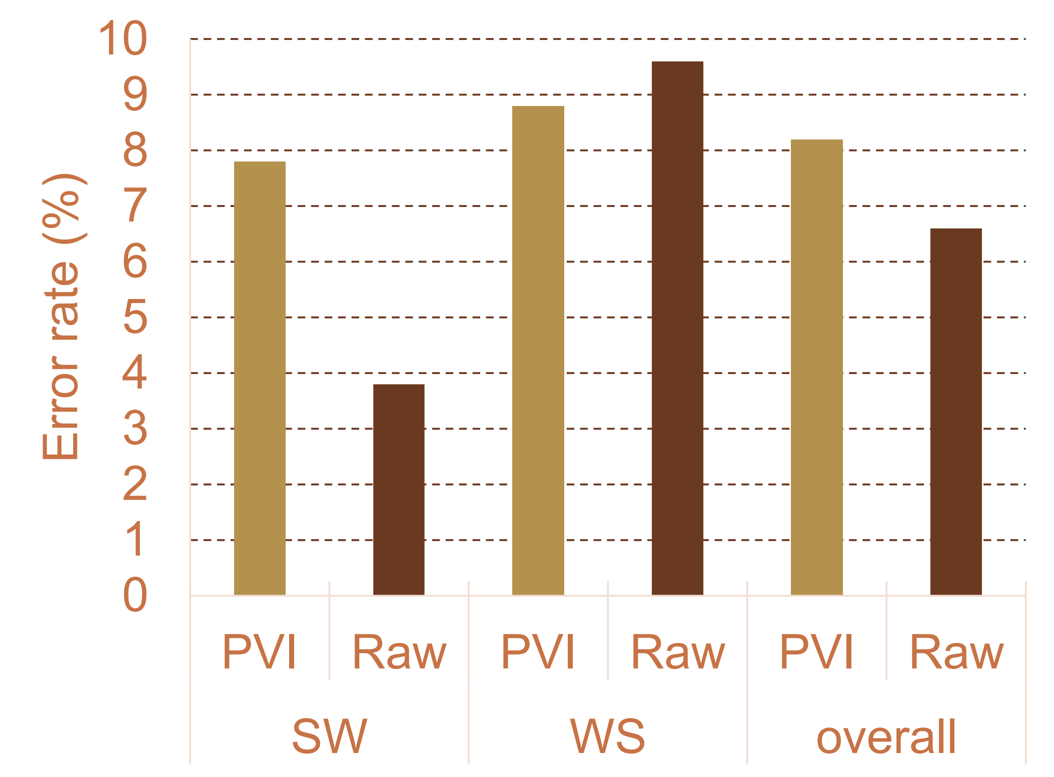
Best ER 12% @ 4 layers/500 units

Raw feature DNN (typically development corpus)

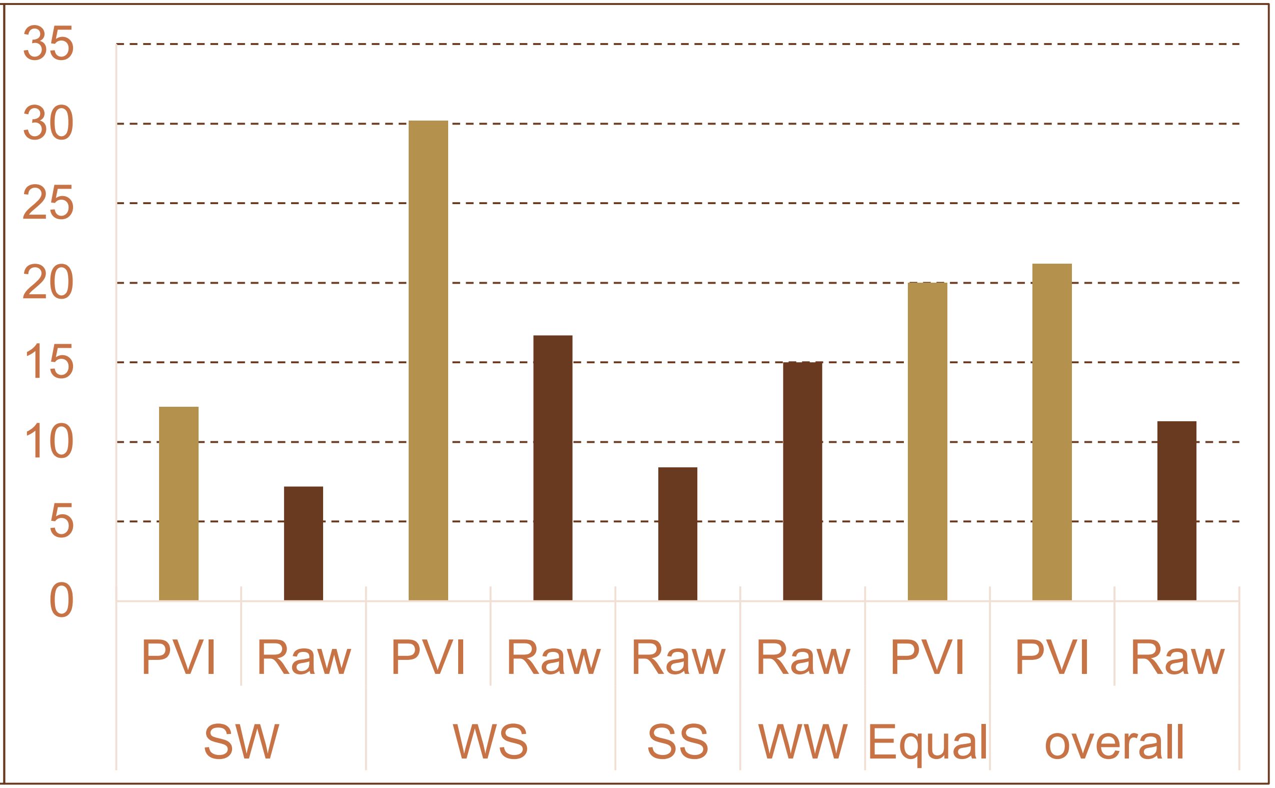
The error rate as a function of number of input frames (N)



Comparison of raw and PVI feature DNN (typically development corpus)



SW/WS



SW/WS/SS/WW

Disordered speech

- System tested against disordered speech which contains only SW/WS patterns.
- The Error rate was:

SW	WS	Overall
27%	25%	26.6%

- The degradation in performance can be explained by the articulation errors that leads to inaccurate phone alignment.
- The perceptual assessment of the disorder speech is inconsistence.
- The inter-rater reliability between two therapists marking lexical stress was 98% for typically developing children and dropped down to 82% for children with CAS.

Conclusions

- We have presented a DNN classifier to detect bisyllabic lexical stress patterns in multi-syllabic English words.
- The DNN classifier is trained using set of temporal and spectral features extracted from pairs of consecutive syllables.
- The feature set of each pair of consecutive syllables is combined by:
  - concatenating the raw features into one wide vector, or
  - computing a variability index to produce one compact feature vector
- Test results on children speech show that the DNN performs better when trained with raw features, as they provide more information than the abstract PVI values.

THANKS
Q&A