INCORPOLATION OF TONE MODELS INTO SPEECH RECOGNITION FOR DYSARTHRIC SPEECH

Dr.Nuttakorn Thubthong
Department of Physics, Faculty of Science, Chulalongkorn University

Mr.Prakasith Kayasith
Asistive Technology Center, NECTEC

Adenda

- Background
- Speech corpora
- Baseline system
- Tone modeling
- Conclusion

What is Dysarthria?

- A neurological motor speech impairment
  - characterised by slow, weak, imprecise and/or uncoordinated movements of the speech musculature.
- Speech is often difficult to understand (unintelligible) and variable (inconsistent)
- Frequently associated with other physical disabilities
- 170/100 000

Intelligibility and Consistency

- ‘Normal’ speech will be almost 100% understandable to listener (*intelligibility*) and with few articulatory differences over time (*consistency*).

- ‘Severe’ dysarthric speech may be completely *unintelligible* to a Unfamiliar Communication Partner and then shows high variability (*inconsistency*) of speech.

  But somehow, for the same speech, to a Familiar Communication Partner it shows some *consistency of key elements* which will make it more *intelligible* to the familiar listener.

Benefits of ASR for dysarthria

- to convert human speech signals into effective actions (ECU)
- to use ASR as an interface to type or send signals to a speech synthesizer that would translate difficult to understand dysarthric speech into a more recognizable form

- ASR is concerned with *consistency of key elements*

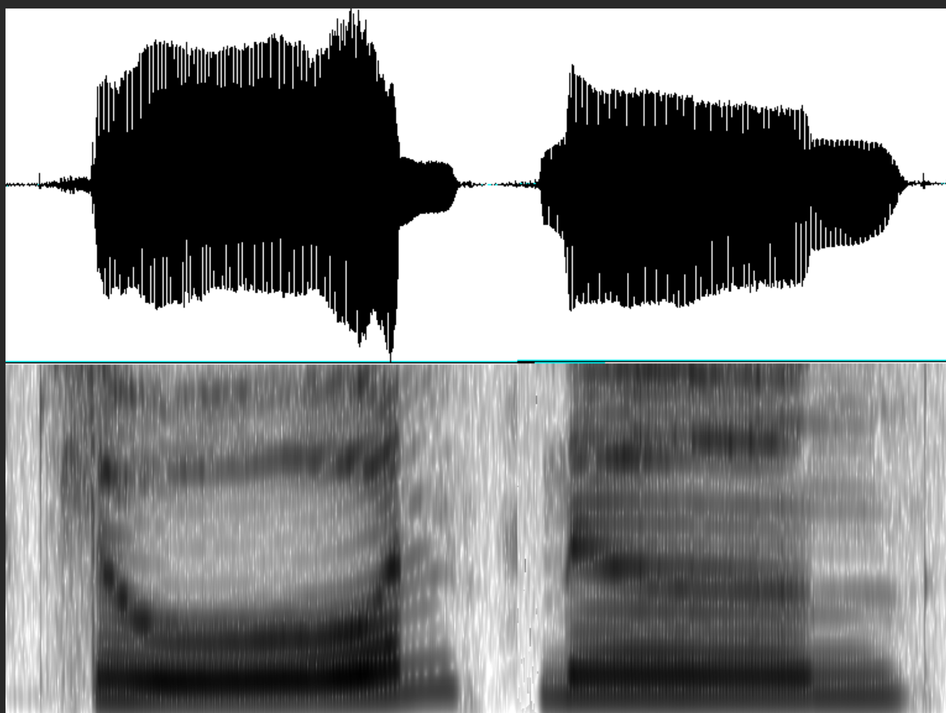
Commercial SRS Tests

	Preselected	Novel
Microsoft Dictation (Microsoft Corporation)	52.23% (87.96%)	52.93% (78.85%)
Dragon Naturally Speaking 3.0 (Dragon Systems Inc.)	67.40% (92.04%)	61.96% (87.72%)
VoicePad Platinum (Kurzweil Education Systems Inc.)	46.99% (85.82%)	56.75% (82.36%)

Data from ISAA (International Society for Augmentative and Alternative Communication, Sept. 2000)

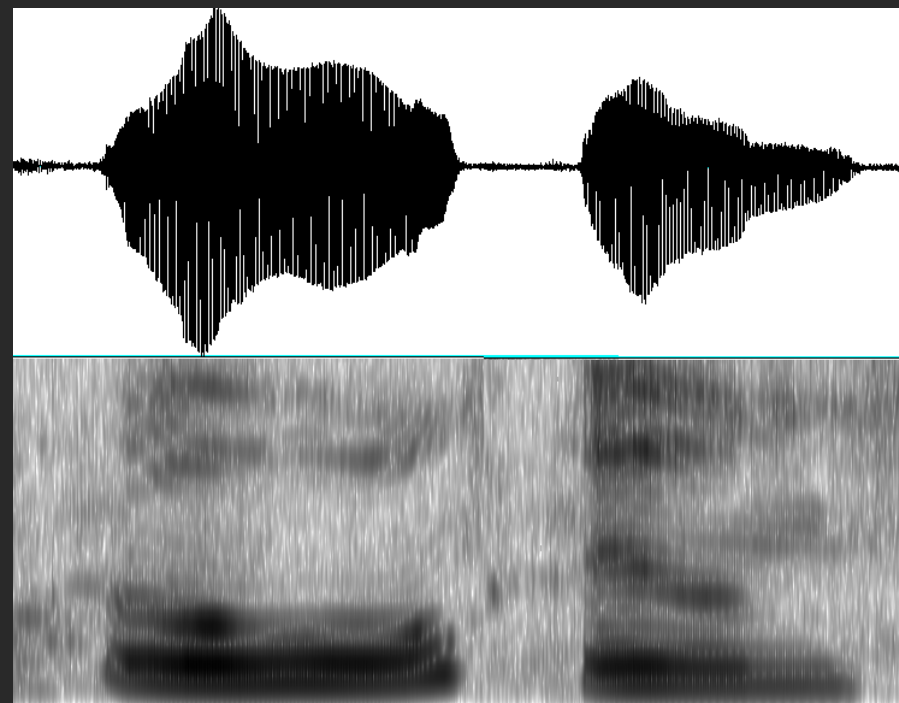
Consonant Error

Normal



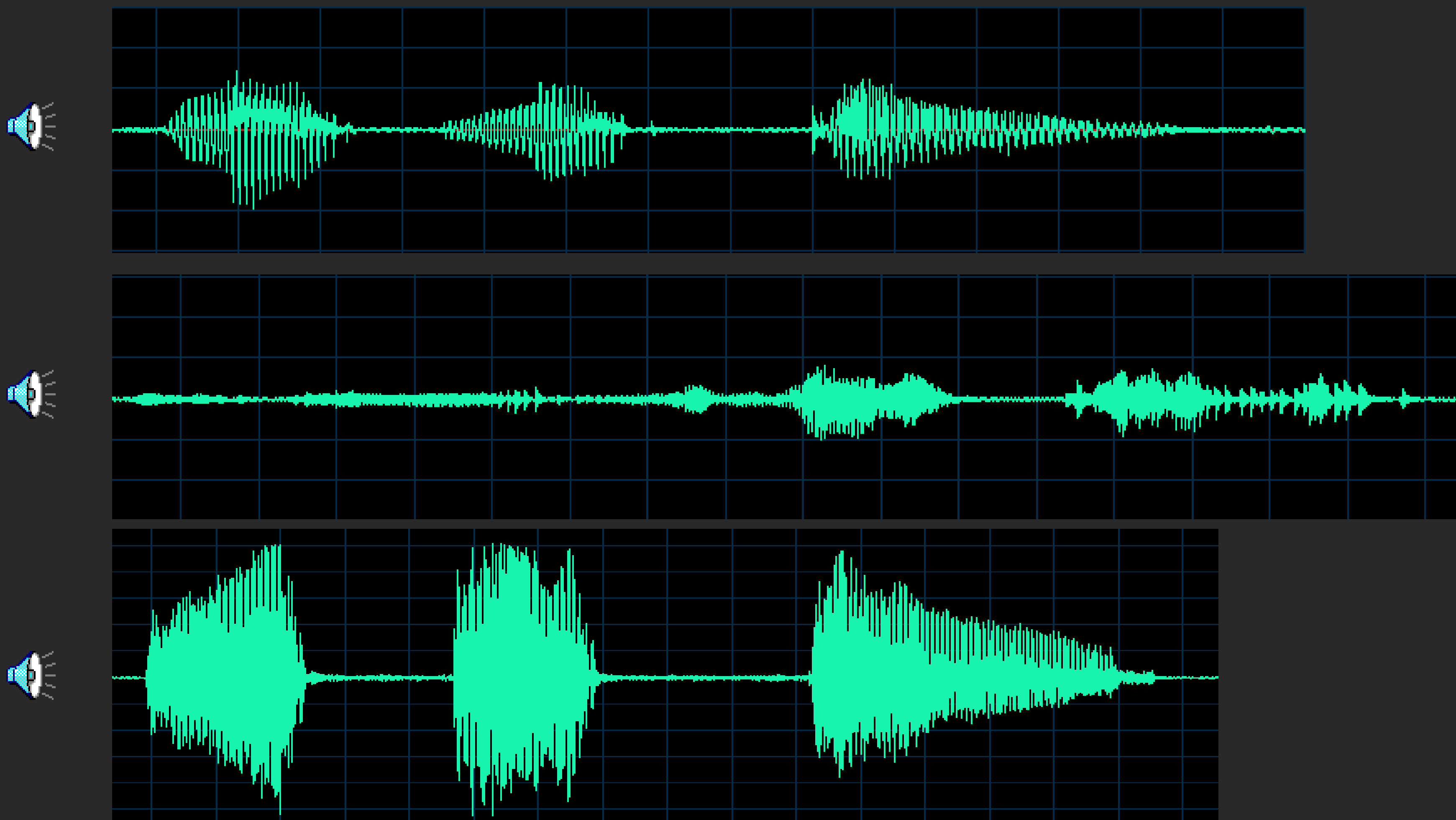
/suung4 nvng1/

Dysarthria

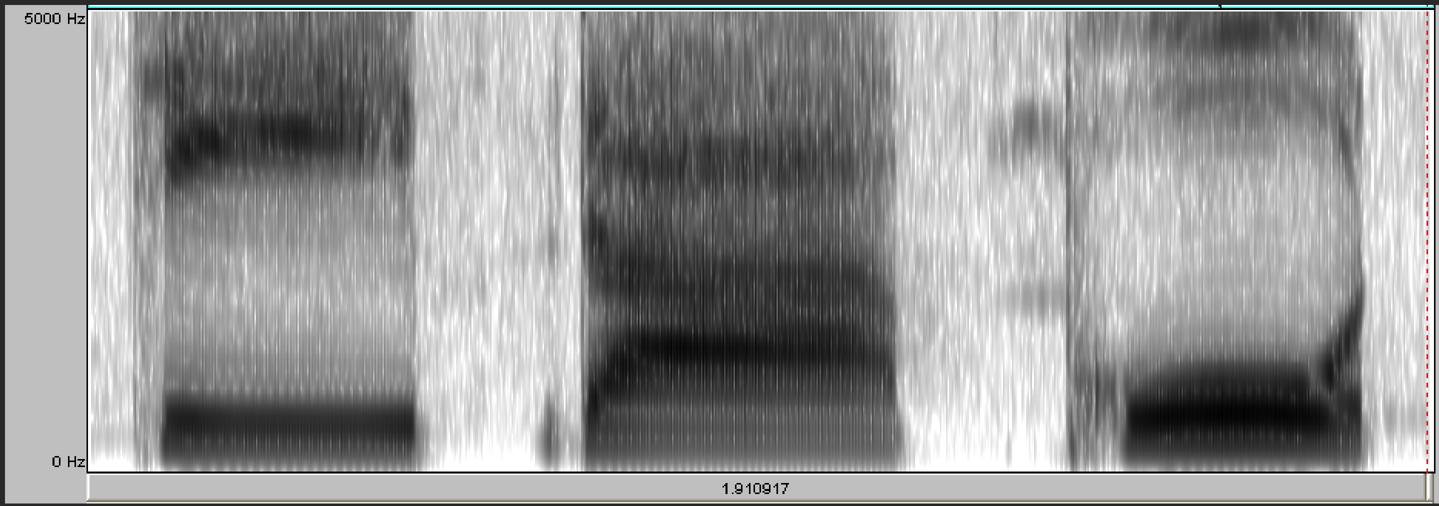


/suung4 nvng1/

Word “มะละกอ” produced by one Normal and two Dysarthria speakers

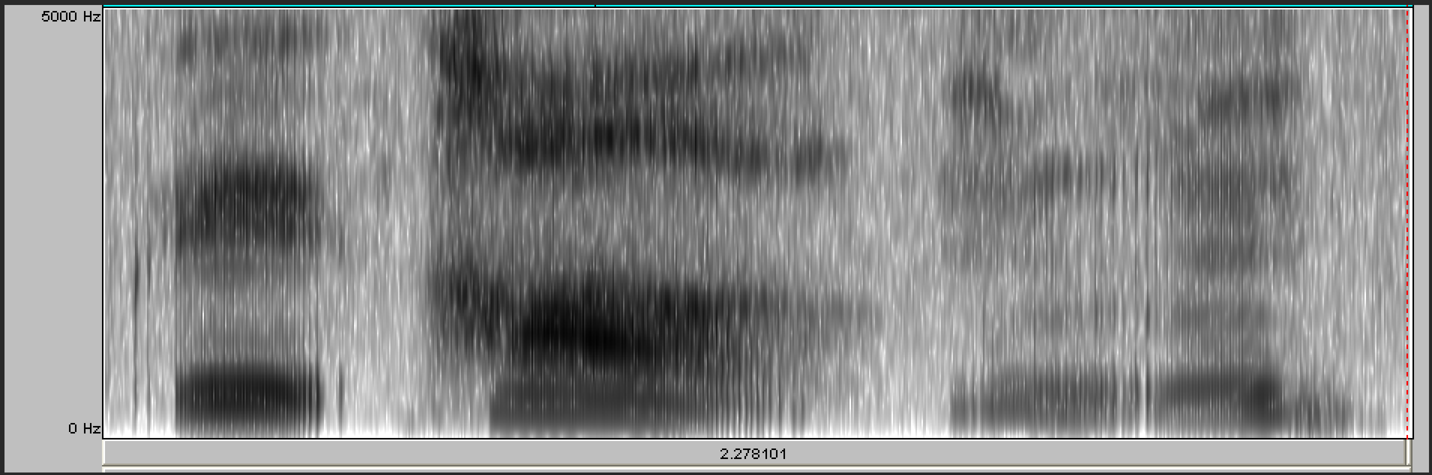


Spectrogram



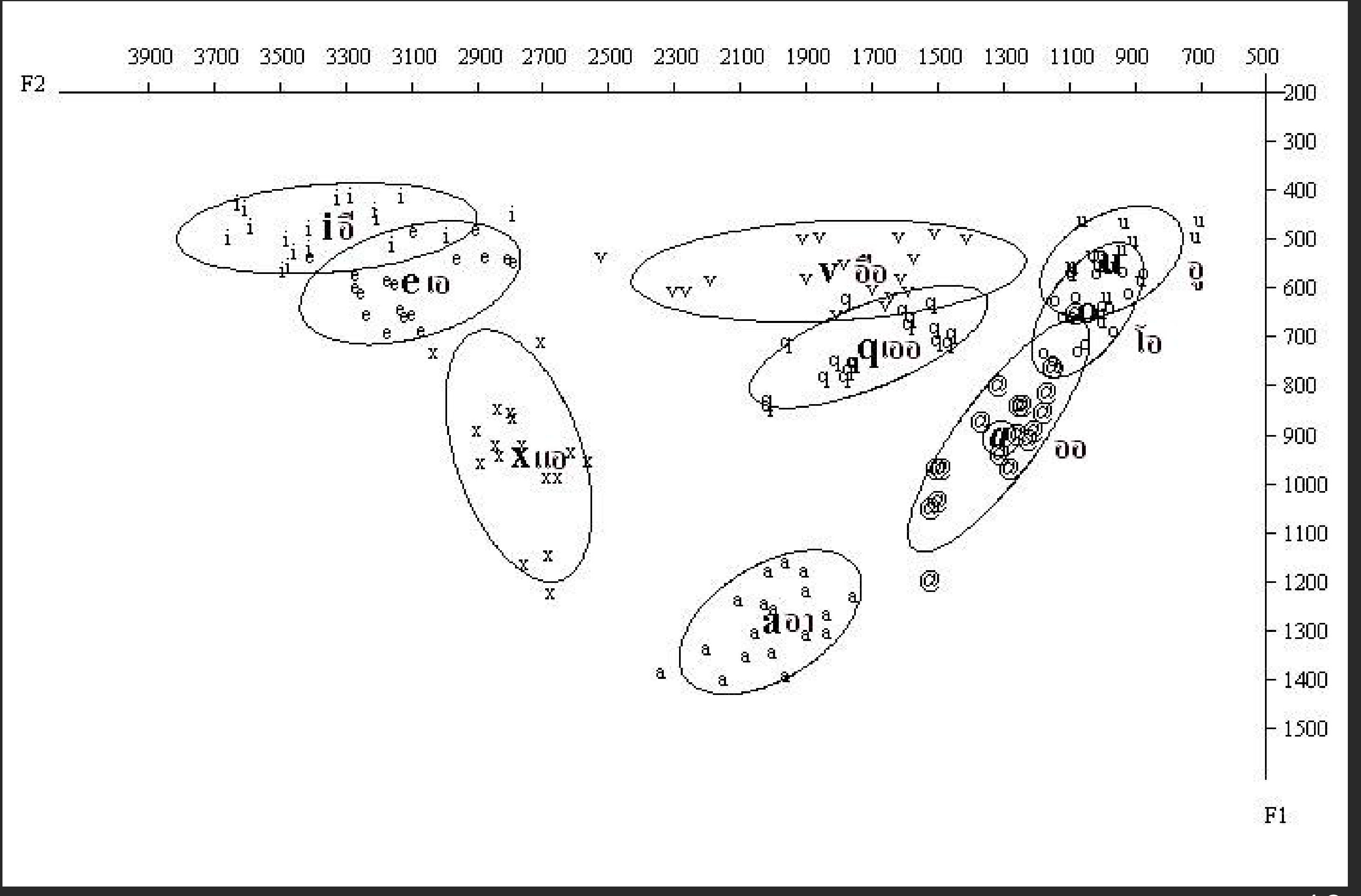
Normal

\ciip1 daap1 phuut2\

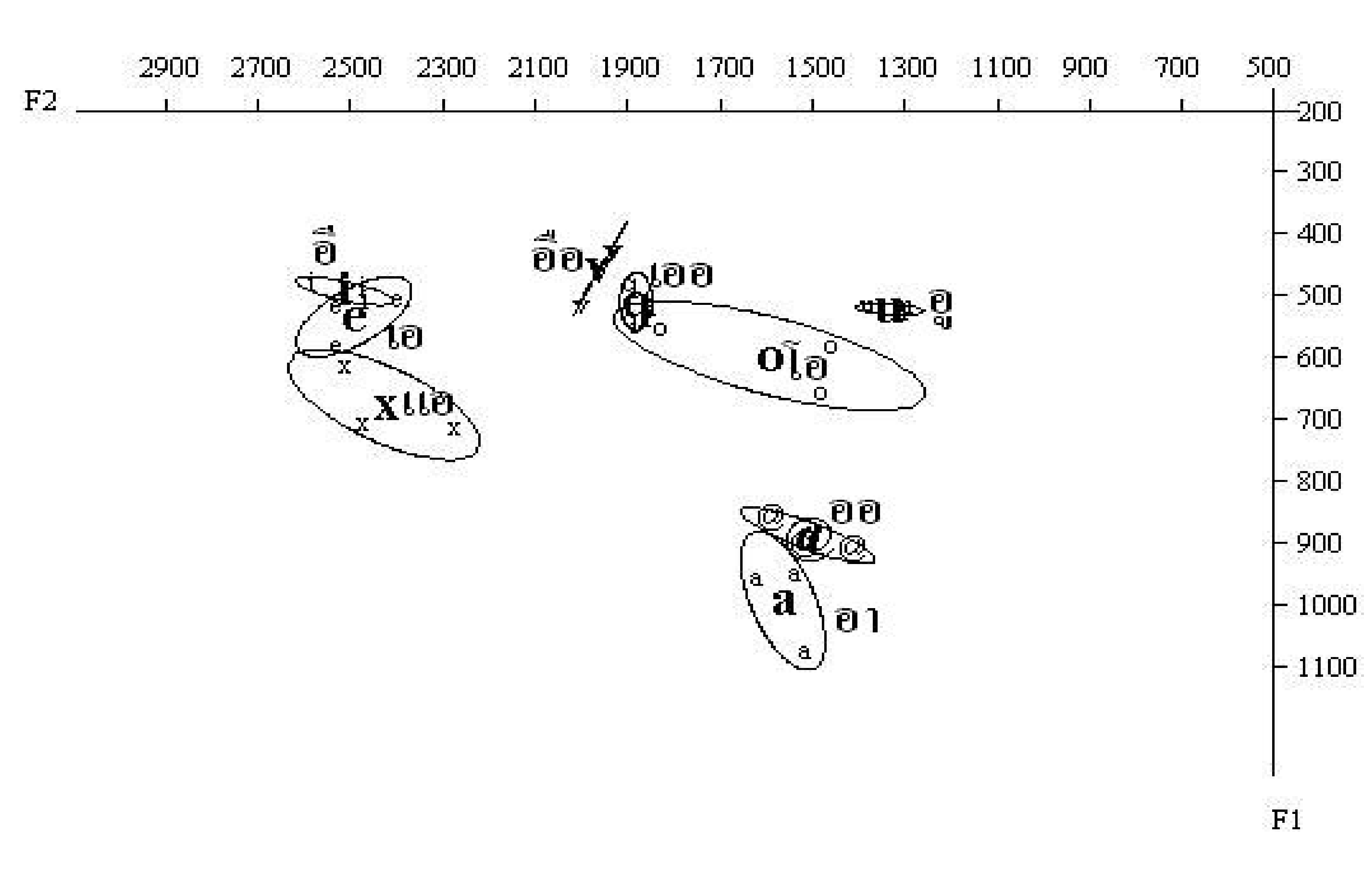


Dysarthria

Vowel areas of six normal speakers

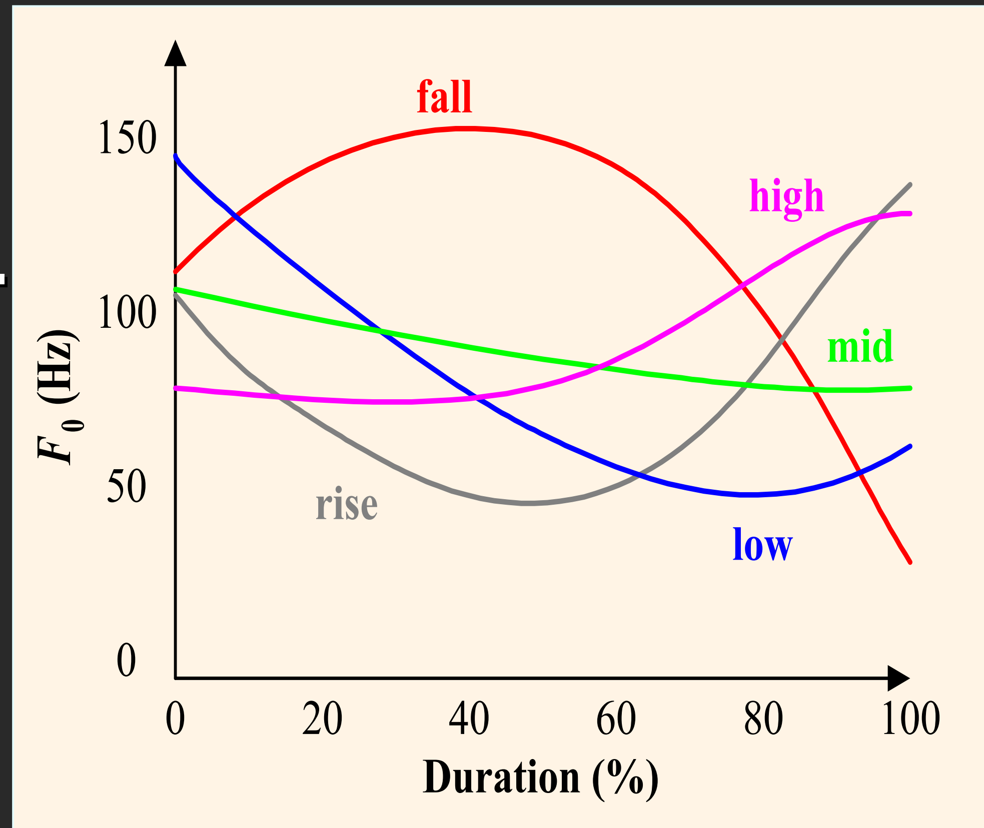


Vowel areas of one dysarthria speaker



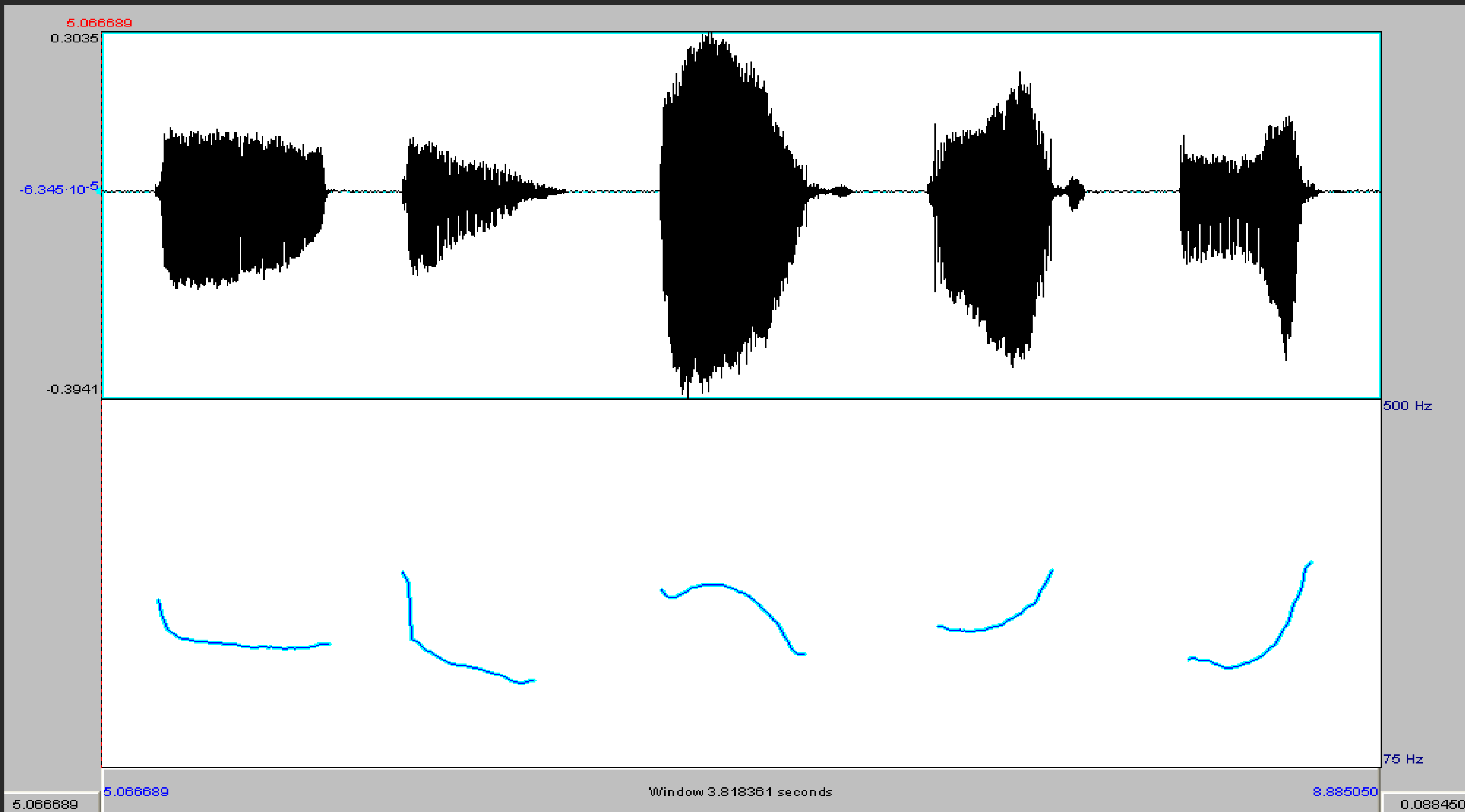
Thai Tones

- Thai is a tone language.
- There are 5 lexical tones.
  - mid สามัญ
  - low เอก
  - fall โท
  - high ตรี
  - rise จัตวา



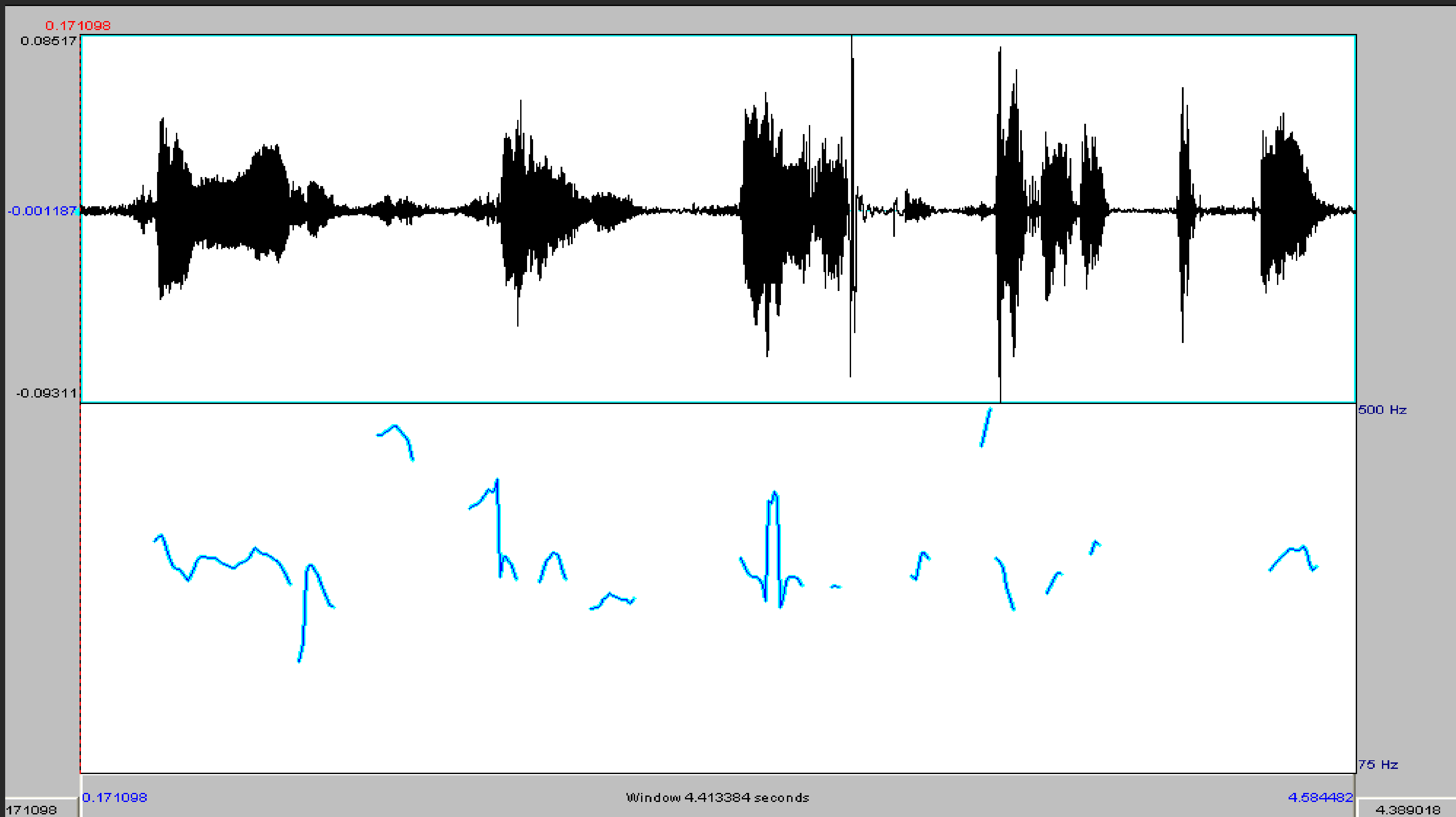
The identification of a Thai tone relies on the shape of ***fundamental frequency*** (***F_0***) contour.

F_0 of Normal Speaker



/paa0 paa1 paa2 paa3 paa4/

F_0 of Dysarthric Speaker



/paa0 paa1 paa2 paa3 paa4/

Speech Corpora

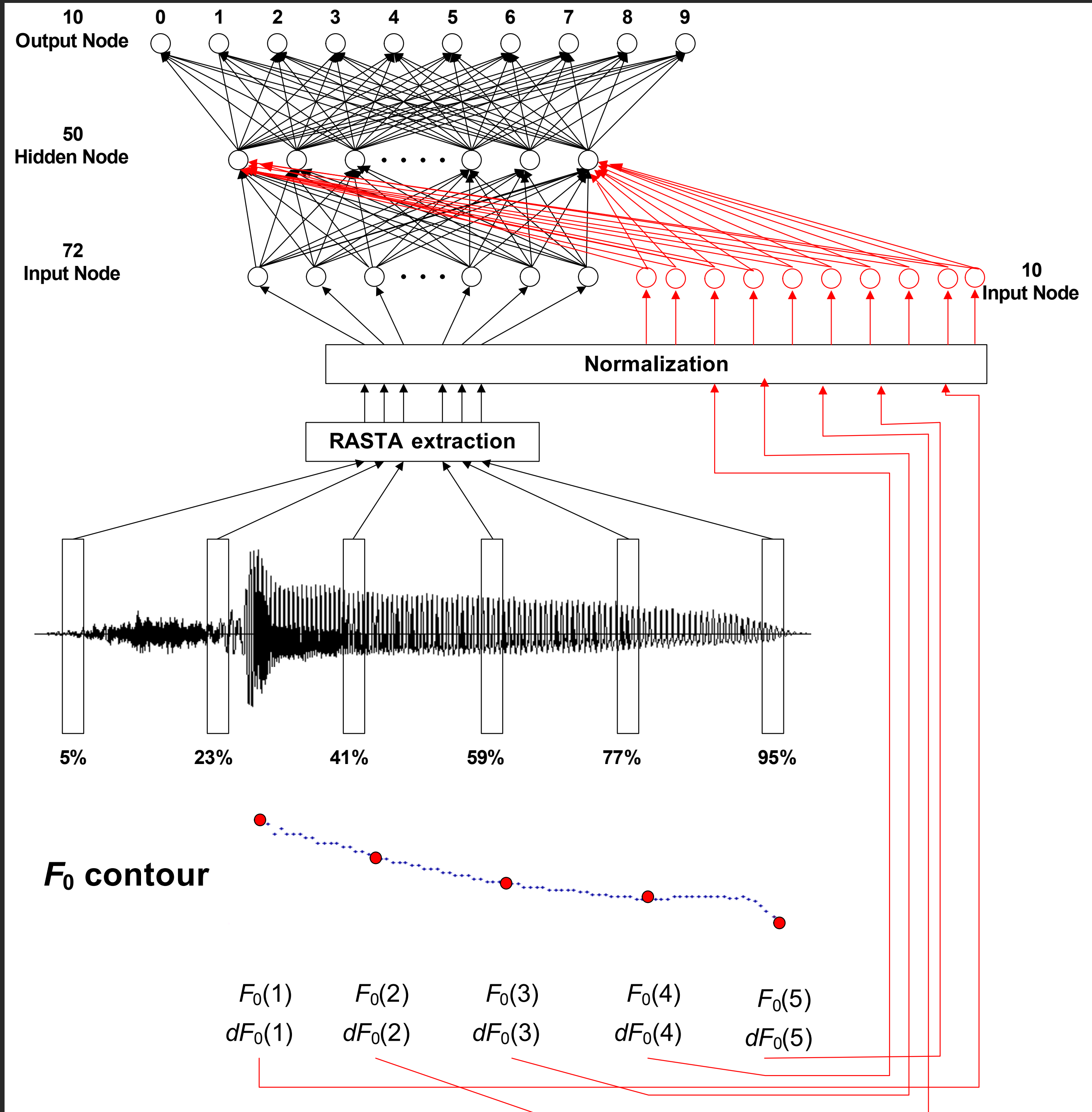
- Three speech corpora:
  - **Digits**: 10 words
  - **Adverbs**: 10 words
  - **Verbs**: 16 words
- 4 normal speakers (6 - 8 years)
- 4 dysarthric speakers (7 - 13 years)
- read all three sets for five trials
- 16-bit, 16 kHz sampling rate.

Baseline system

- 12th RASTA, 25 ms frame size
- Three-layer Feedforward Neural Network
  - 180 input nodes
  - 50-50-100 hidden nodes for digit, adverb and verb sets
  - 10-10-16 for digit, adverb and verb sets
- 5-Fold Cross-Validation
- Trained by the error back-propagation algorithm

Tone Models

- Using Average Magnitude Different Function (AMDF) algorithm
- Normalized by transforming the Hertz values to a z-score using mean and standard deviation of each speaker
- Smoothed using the 3^{rd} order polynomial regression
- Use five F0's and their derivatives at 0, 25, 50, 75, and 100% as tone models


10
Output Node
0
1
2
3
4
5
6
7
8
9
50
Hidden Node
72
Input Node
10
Input Node
Normalization
RASTA extraction
5%
23%
41%
59%
77%
95%
F_0 contour
$F_0(1)$
$dF_0(1)$
$F_0(2)$
$dF_0(2)$
$F_0(3)$
$dF_0(3)$
$F_0(4)$
$dF_0(4)$
$F_0(5)$
$dF_0(5)$

Experiment results for normal speaker

	DIGIT (10)		ADVERB (10)		VERB (16)	
	Baseline	+TONE	Baseline	+TONE	Baseline	+TONE
M1	100.0	100.0	100.0	100.0	98.7	100.0
M2	100.0	100.0	100.0	100.0	97.5	98.7
F1	96.0	98.0	98.0	100.0	97.5	98.7
F2	100.0	100.0	98.0	98.0	95.0	95.0
Avg	**99.0**	**99.5**	**99.0**	**99.5**	**97.2**	**98.1**

Experiment results for dysarthric speaker

	DIGIT (10)		ADVERB (10)		VERB (16)	
	Baseline	+TONE	Baseline	+TONE	Baseline	+TONE
DM1	76.0	80.0	80.0	84.0	68.7	71.2
DM2	92.0	94.0	96.0	98.0	92.5	93.7
DF1	98. 0	100.0	94.0	94.0	88.7	92.5
DF2	80.0	86.0	80.0	80.0	-	-
Avg	**86.5**	**90.0**	**87.5**	**89.0**	**83.3**	**85.8**

Conclusion

- ASR is concerned with consistency of key elements
- We demonstrated that incorporating tone models improved the recognition performance for dysarthric speakers.
- Other prosodic information such as duration and stress modeling will be investigated in our future work.