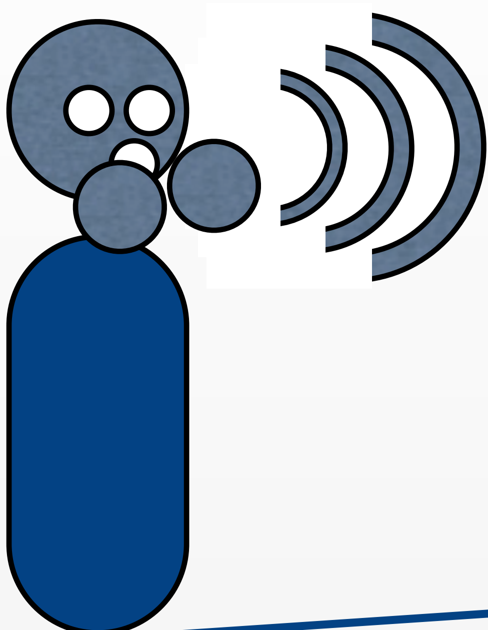
Towards articulatory-based adaptation in recognition of dysarthric speech

Frank Rudzicz
PhD candidate
Department of Computer Science
University of Toronto

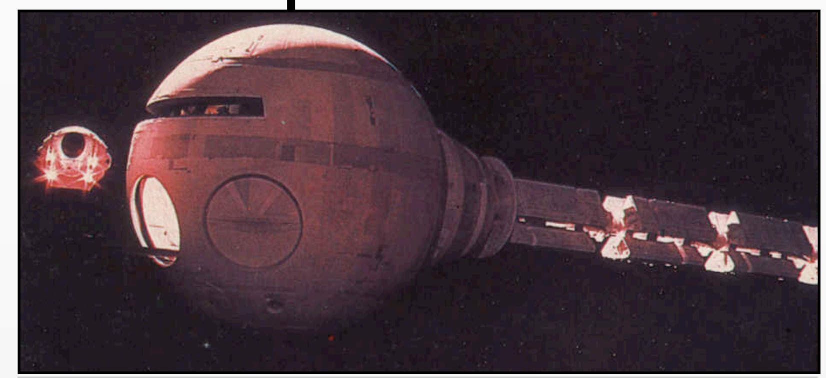
UNIVERSITY of TORONTO

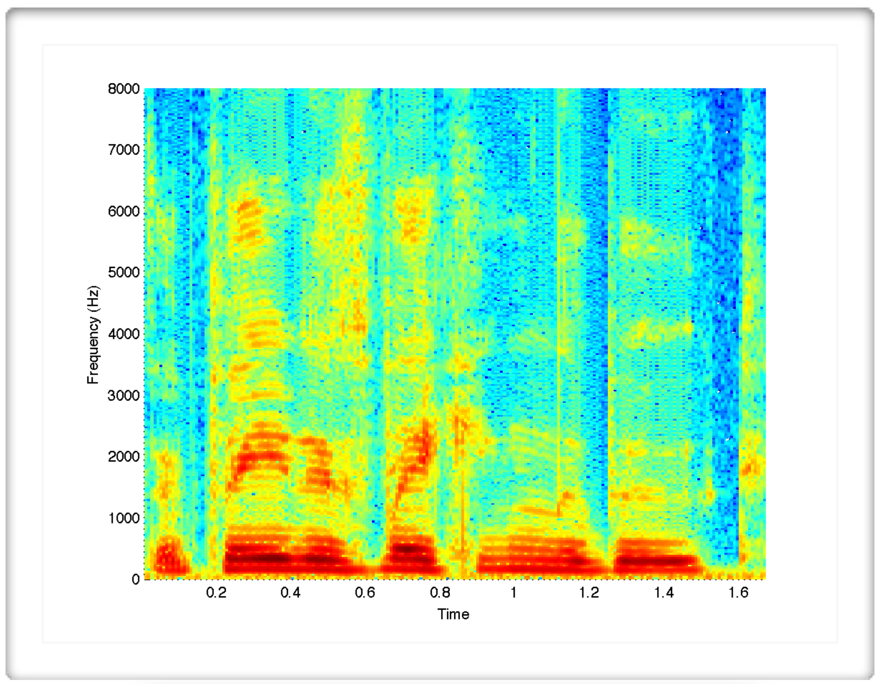
Automatic speech recognition (ASR)



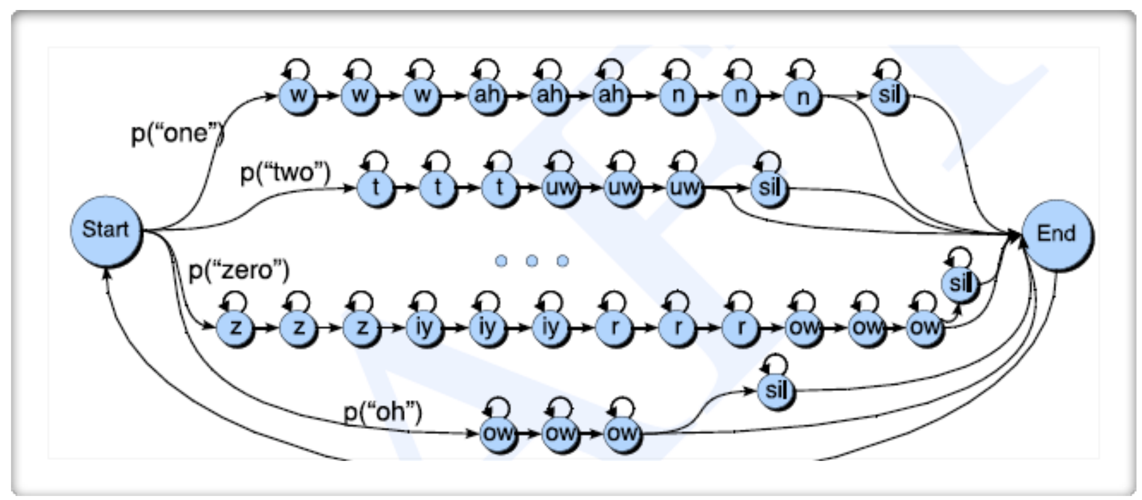
“open the pod bay doors”

open(podBay.doors);





Acoustic model



Language model

Dysarthria

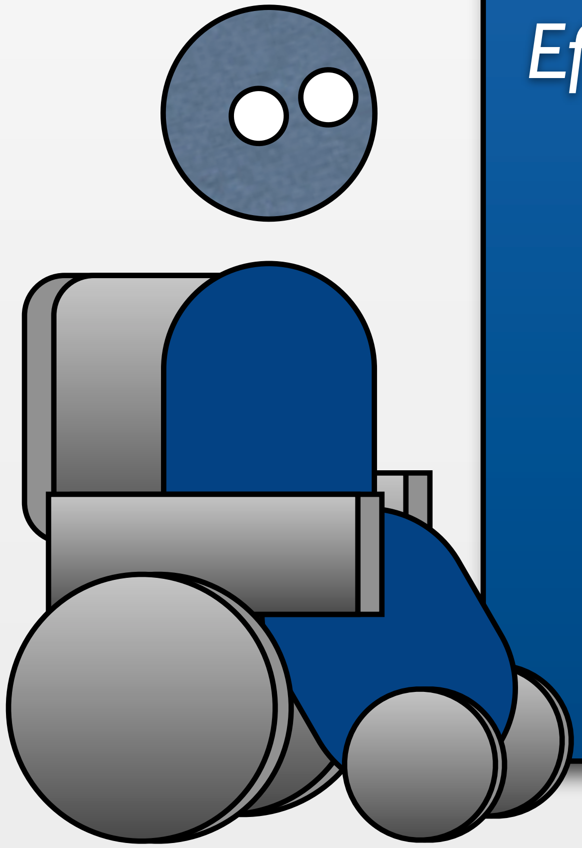
Articulatory disabilities resulting in unintelligible speech.

Causes (e.g.)

- Cerebral palsy.
- Parkinson's disease.
- Amyotrophic lateral sclerosis.

Effects (e.g.)

- Poor respiration, phonation, and resonance.
- Inaccurate timing and speed.
- Imprecise consonants.
- Distorted or indistinguishable vowels.
- Atypical control of volume and pitch.



Dysarthria and speech recognition

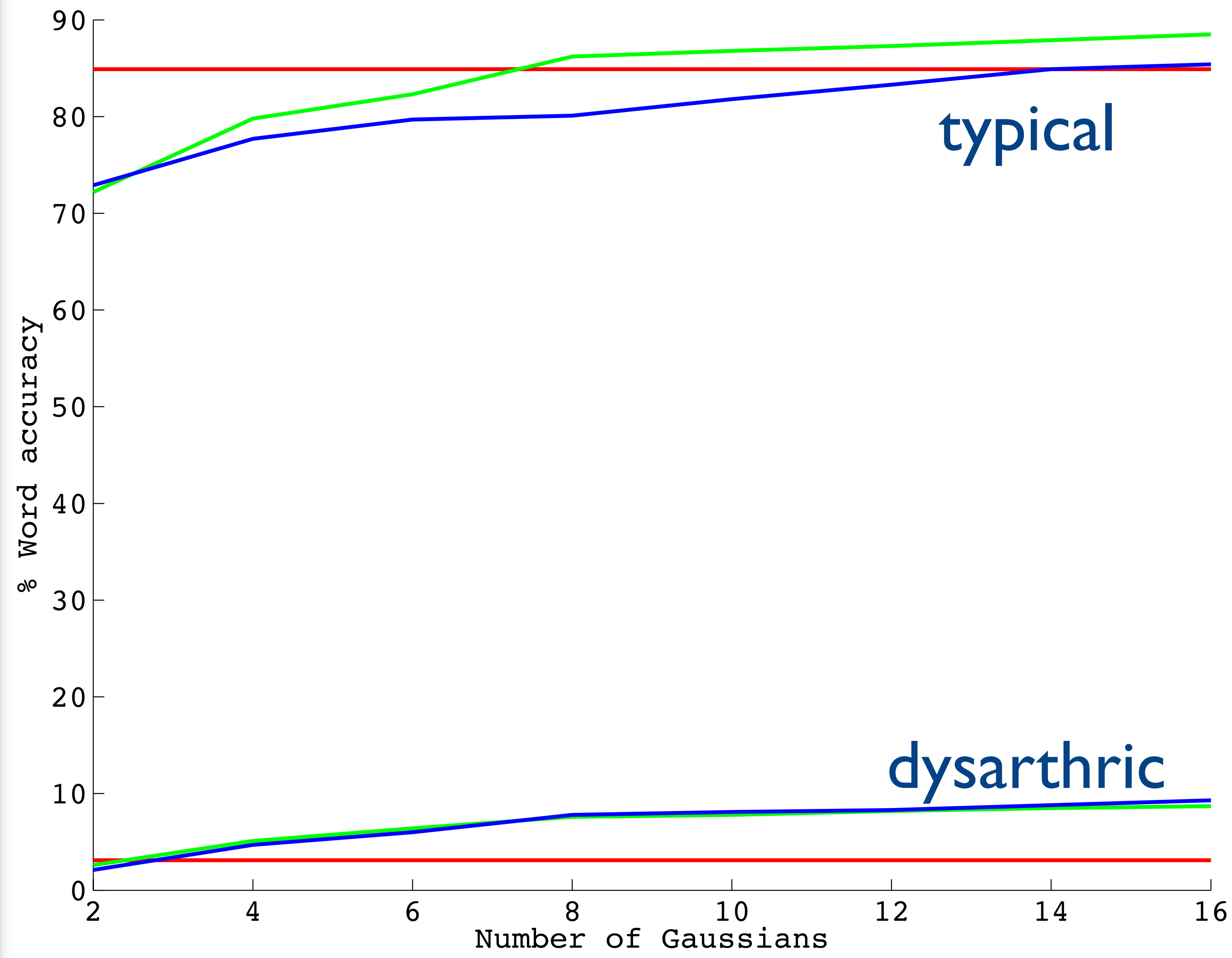
Word-recognition accuracy given traditional ASR (red):

- **84.9%** with typical speech.
- **3.1%** with severe dysarthria.

We can improve accuracy by

- Increasing the complexity of our models (# Gaussians).
- Adapting existing models to new data (green).
- Learning from scratch (blue).

Still, fewer than 10% of words will be recognized. *Why?*



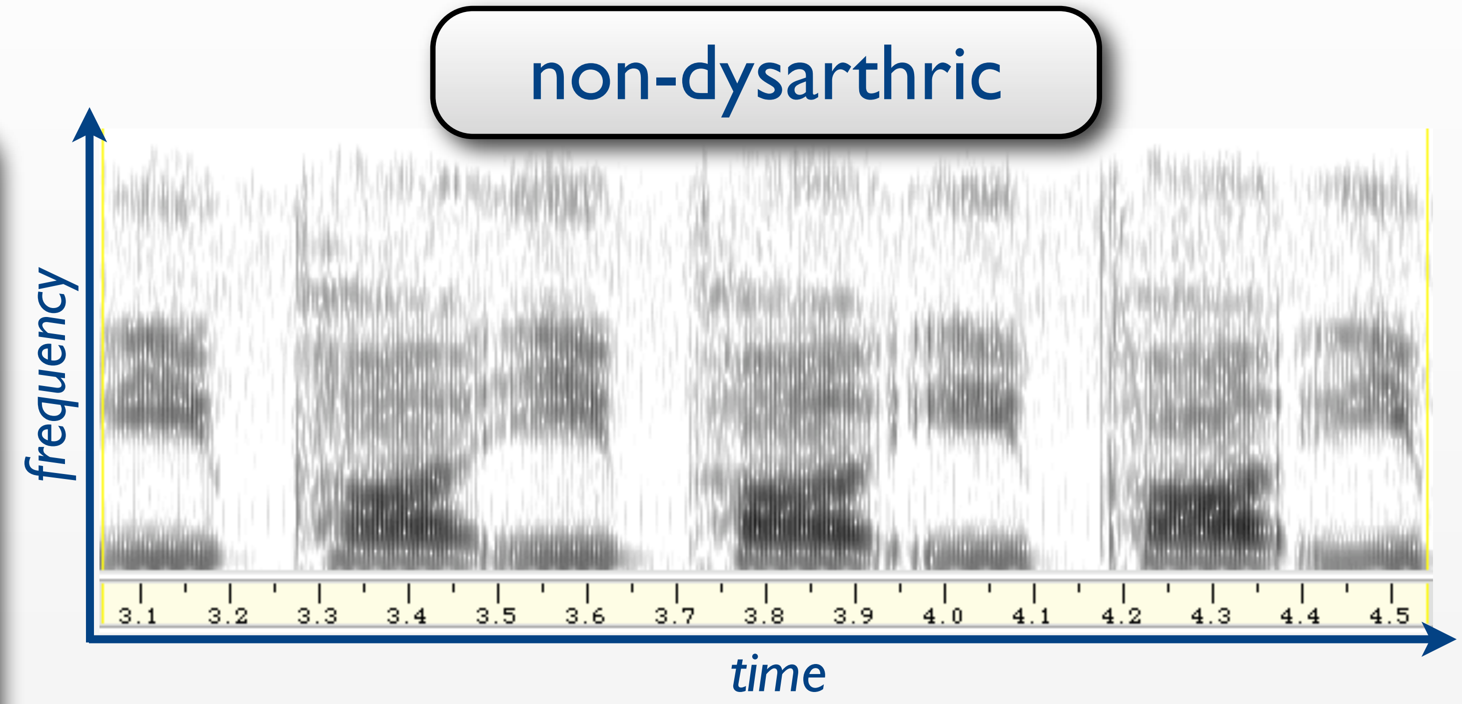
UNIVERSITY of TORONTO

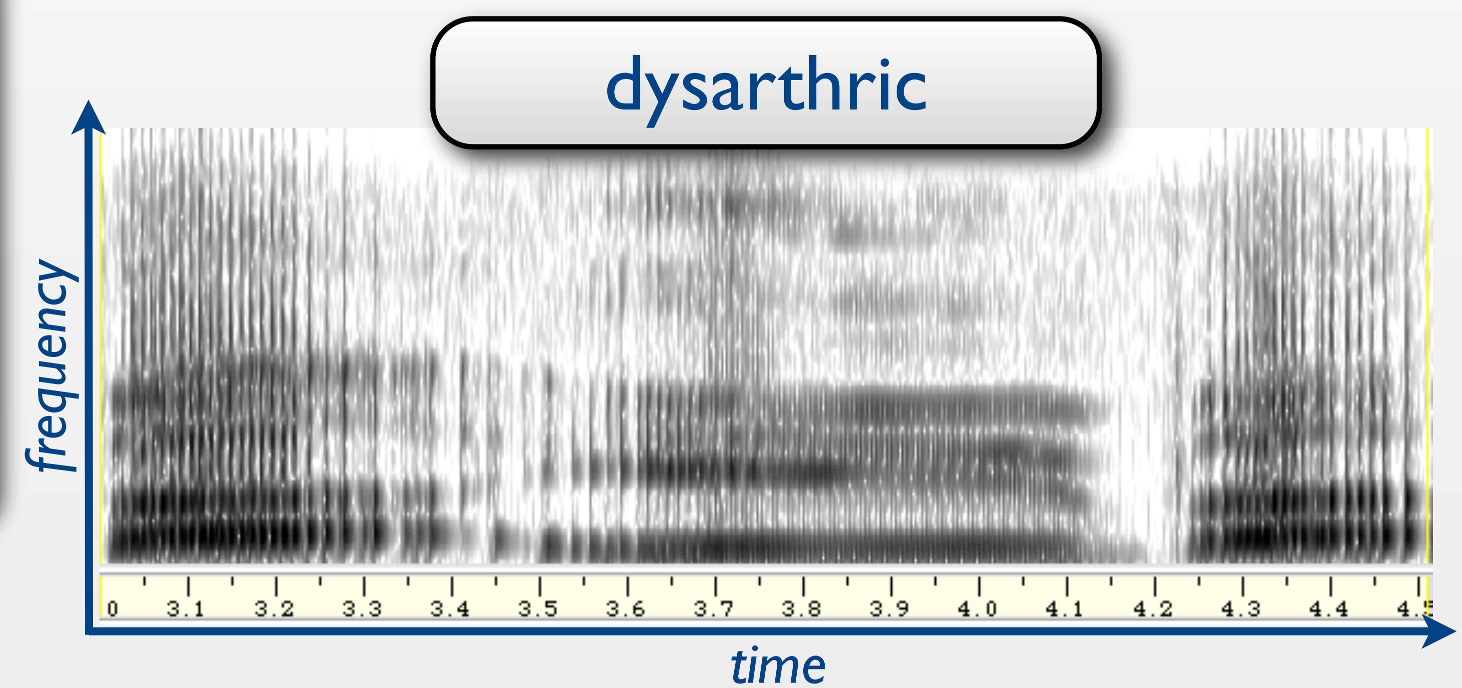
Consistency, disfluency, accuracy

ASR has trouble with dysarthric speech because

- it is inconsistent.
- it has indistinguishable targets.
- it is interrupted by disfluencies (e.g., stutter).

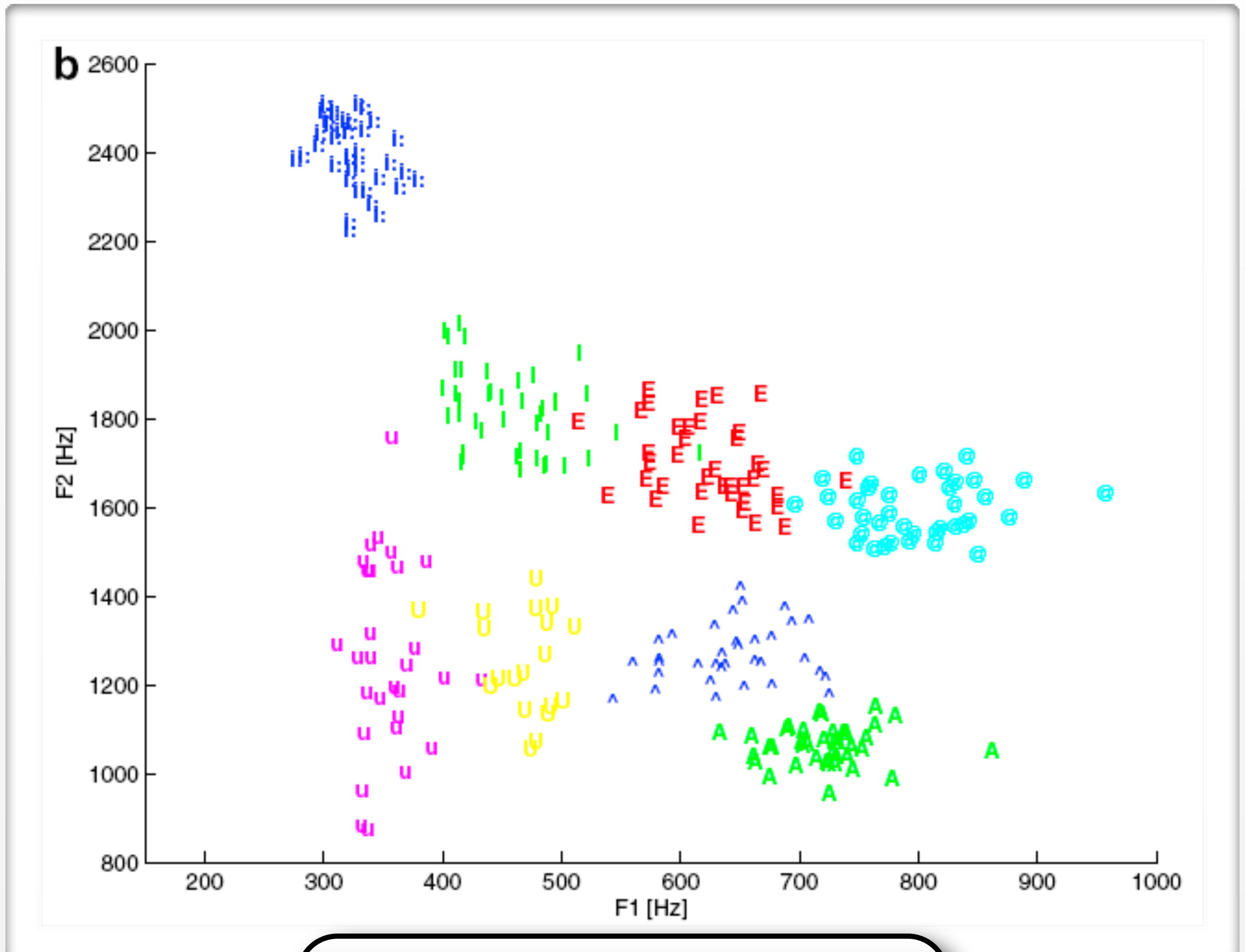
E.g., given repetitions of the sequence /iy p ae/ (right)



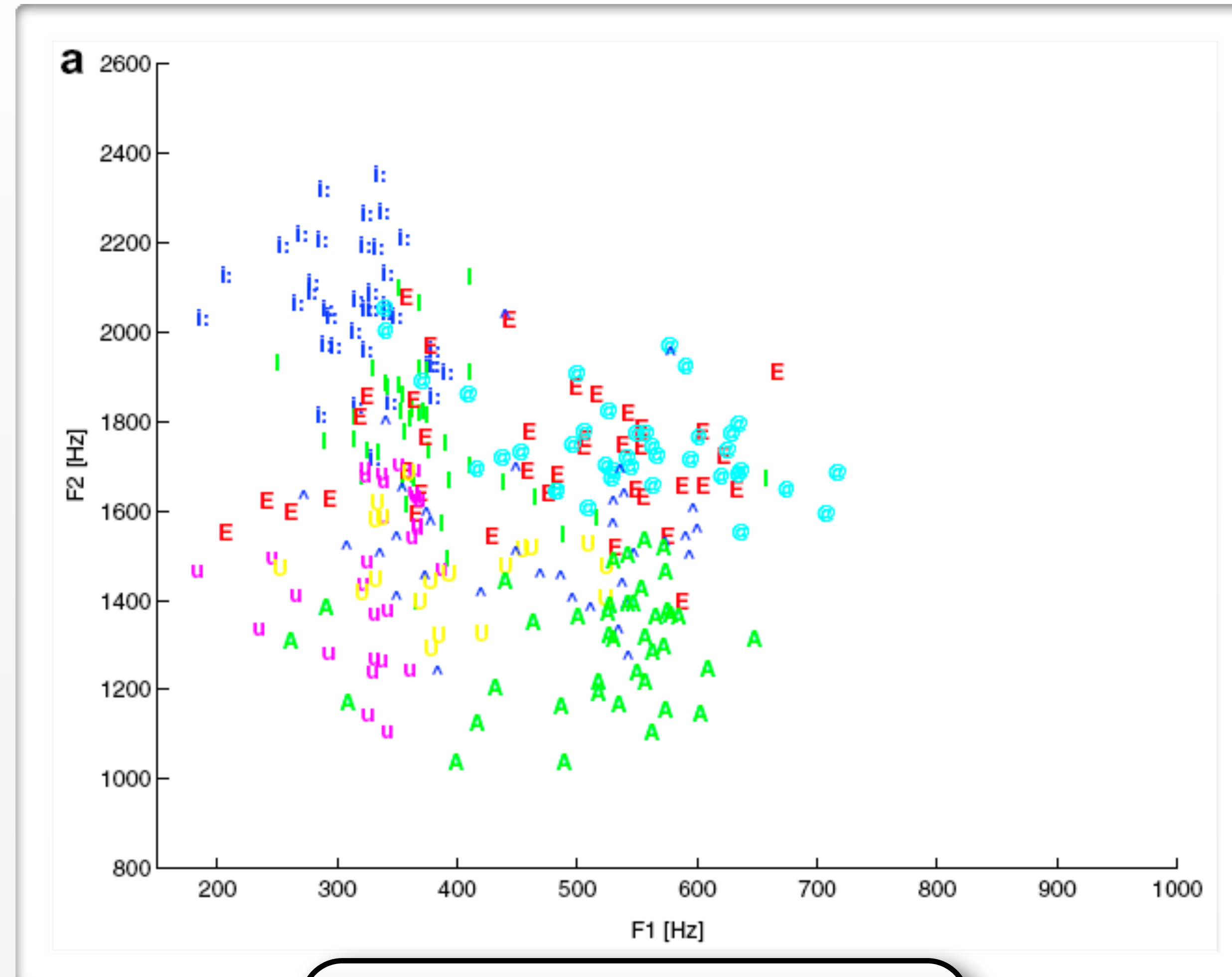


UNIVERSITY of TORONTO

Acoustic distribution of vowels



non-dysarthric



dysarthric

Non-dysarthric speakers can produce very distinct vowels.

It is difficult to correctly associate an observed sound with a cluster if those clusters are not clearly delineated.

Towards articulatory models

Dysarthria is characterized by aberrant neuro-motor signals. The result is atypical articulation.

Can we identify how dysarthric articulation differs from the general public?

Can we use these articulatory models within ASR?

Data collection

"Torgo"

- A database of aligned acoustics and articulation.

Population

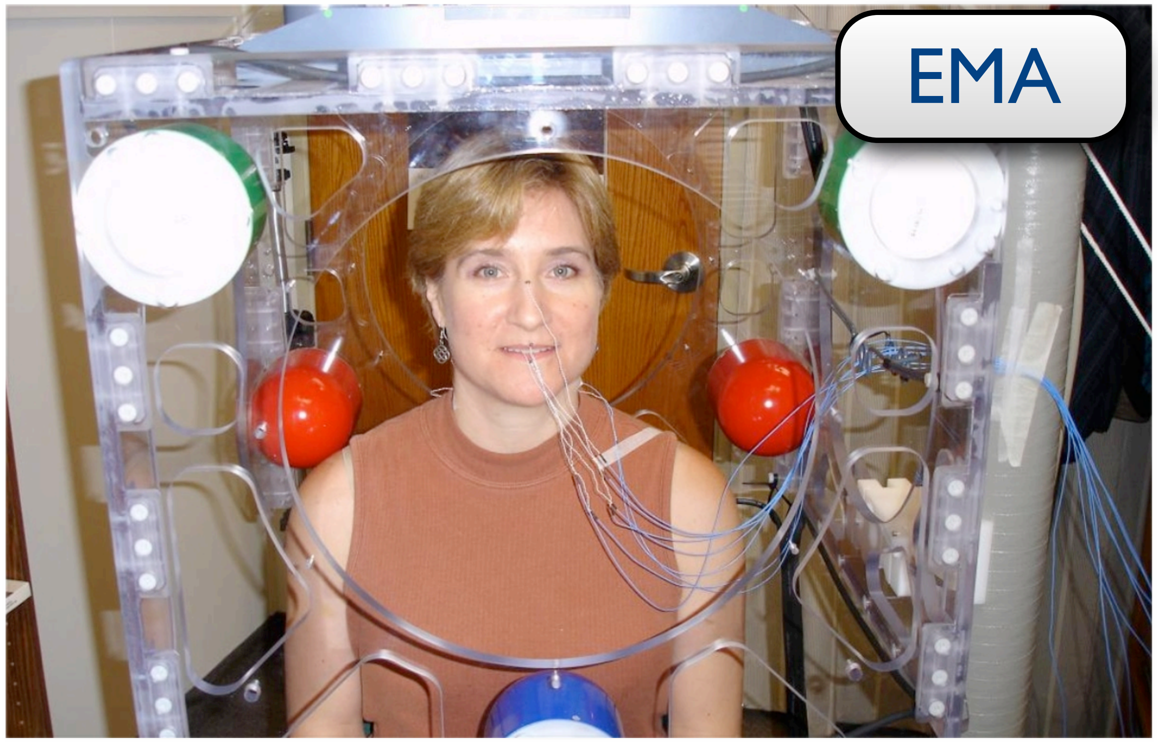
- 7 individuals with CP or ALS,
  + matched controls.

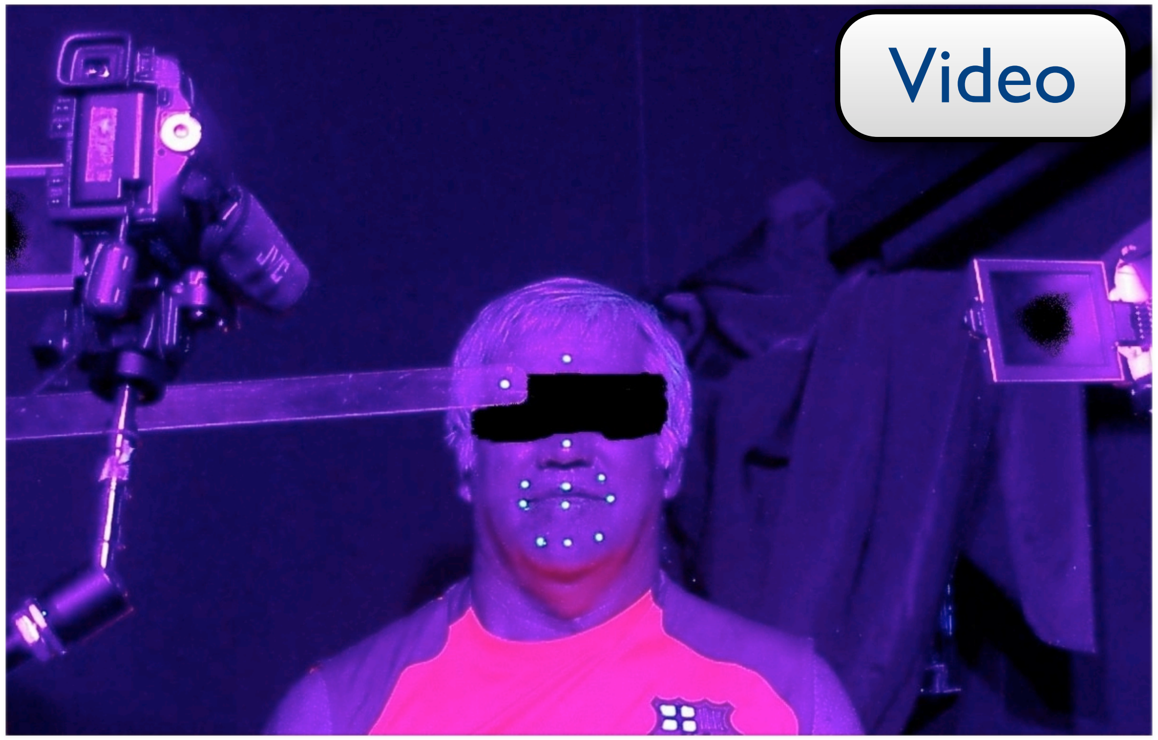
Data

- 3 hrs of sentences and words each.

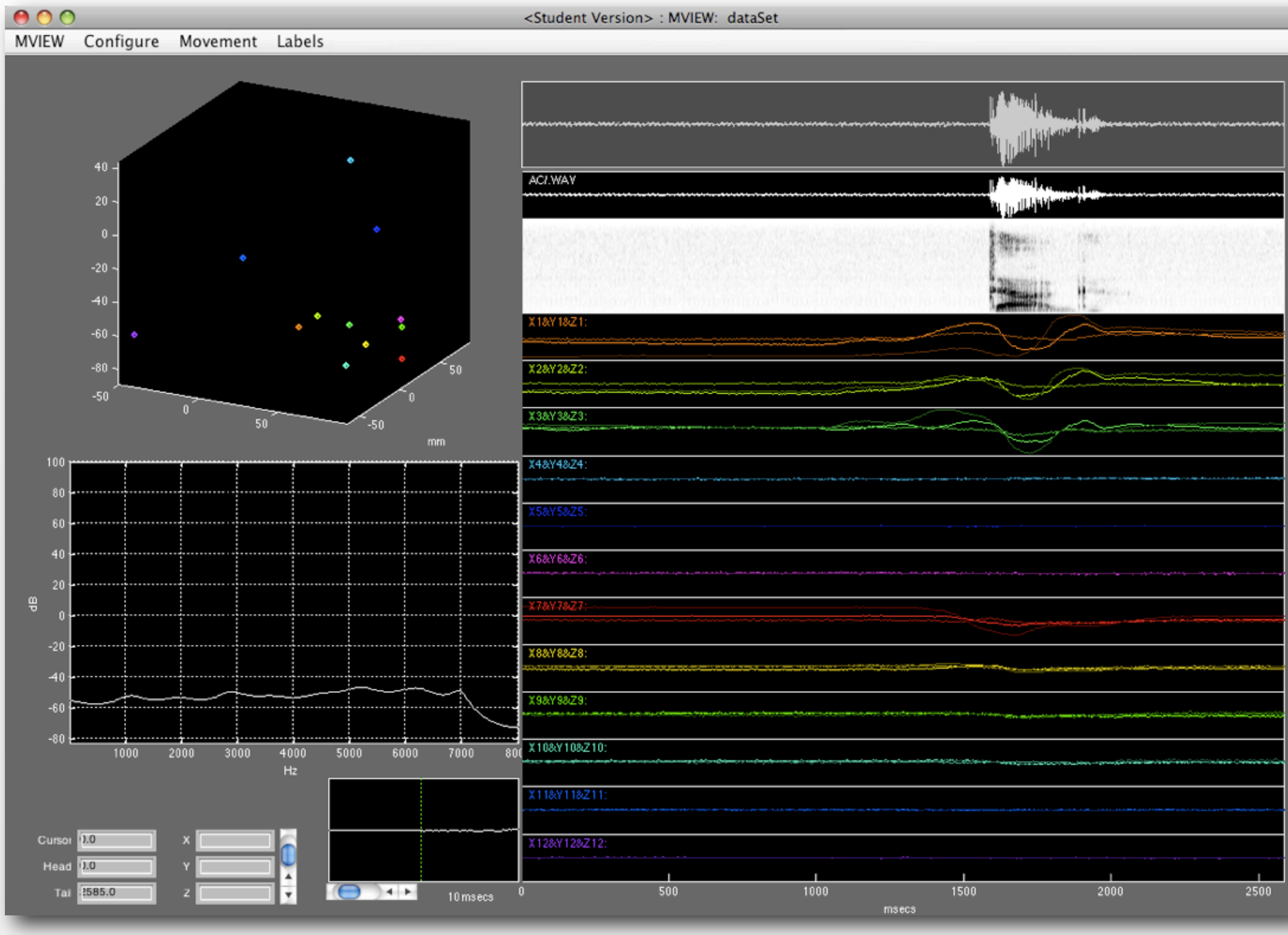
Methods

- Electromagnetic articulography.
- 3D video reconstruction.





Electromagnetic articulography (EMA)



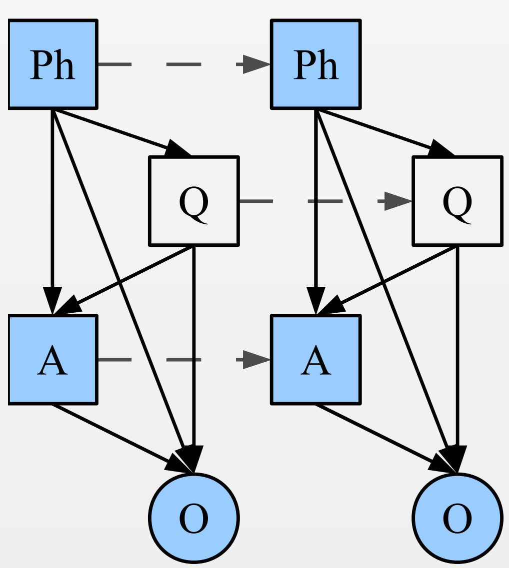
3D position, velocity, and acceleration of coils on tongue and lips are measured at 200Hz.

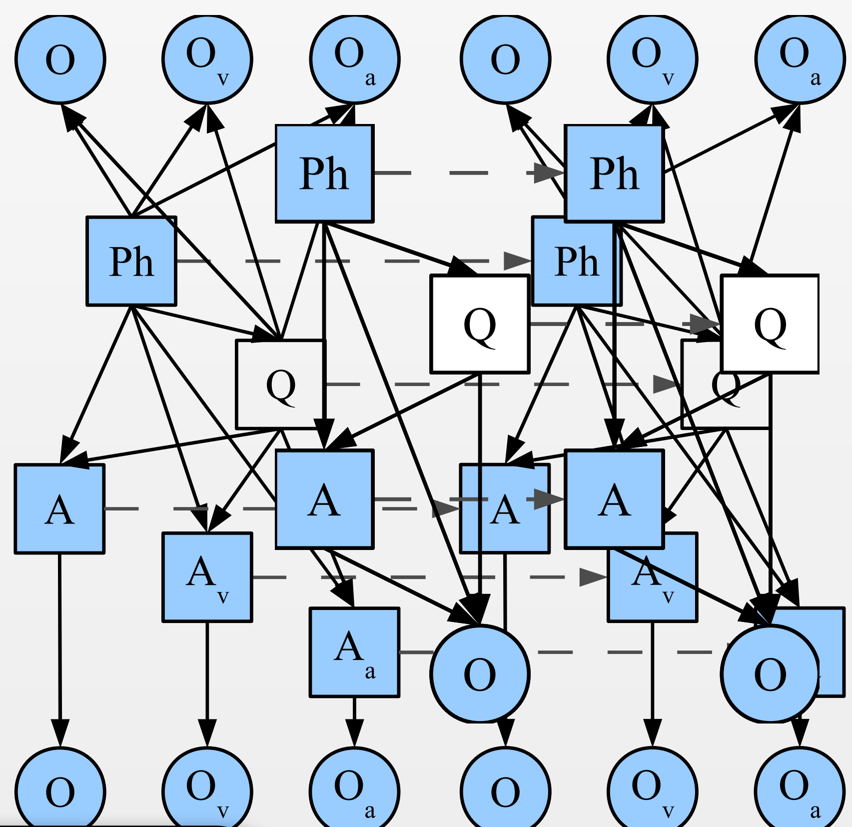
Dynamic patterns of articulation are aligned with their acoustic consequences, and analyzed.

Articulatory models for ASR

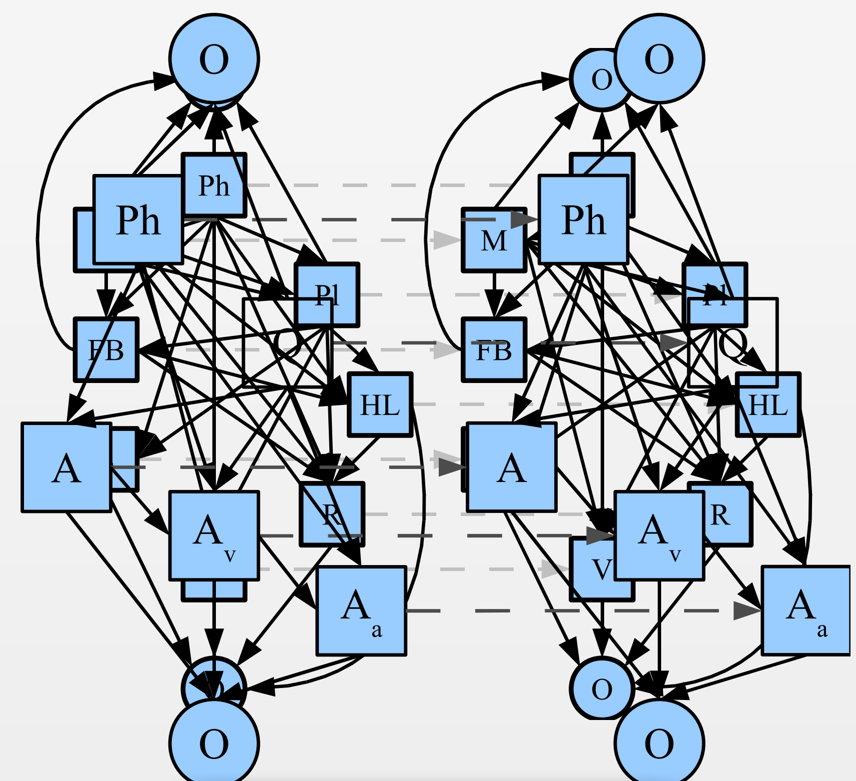
Build models that combine acoustics, articulation, and phonemes.

Dynamic Bayesian networks model the probabilistic behaviour of speech over time, and condition observed acoustics on articulatory sources.





EMA



Phonological

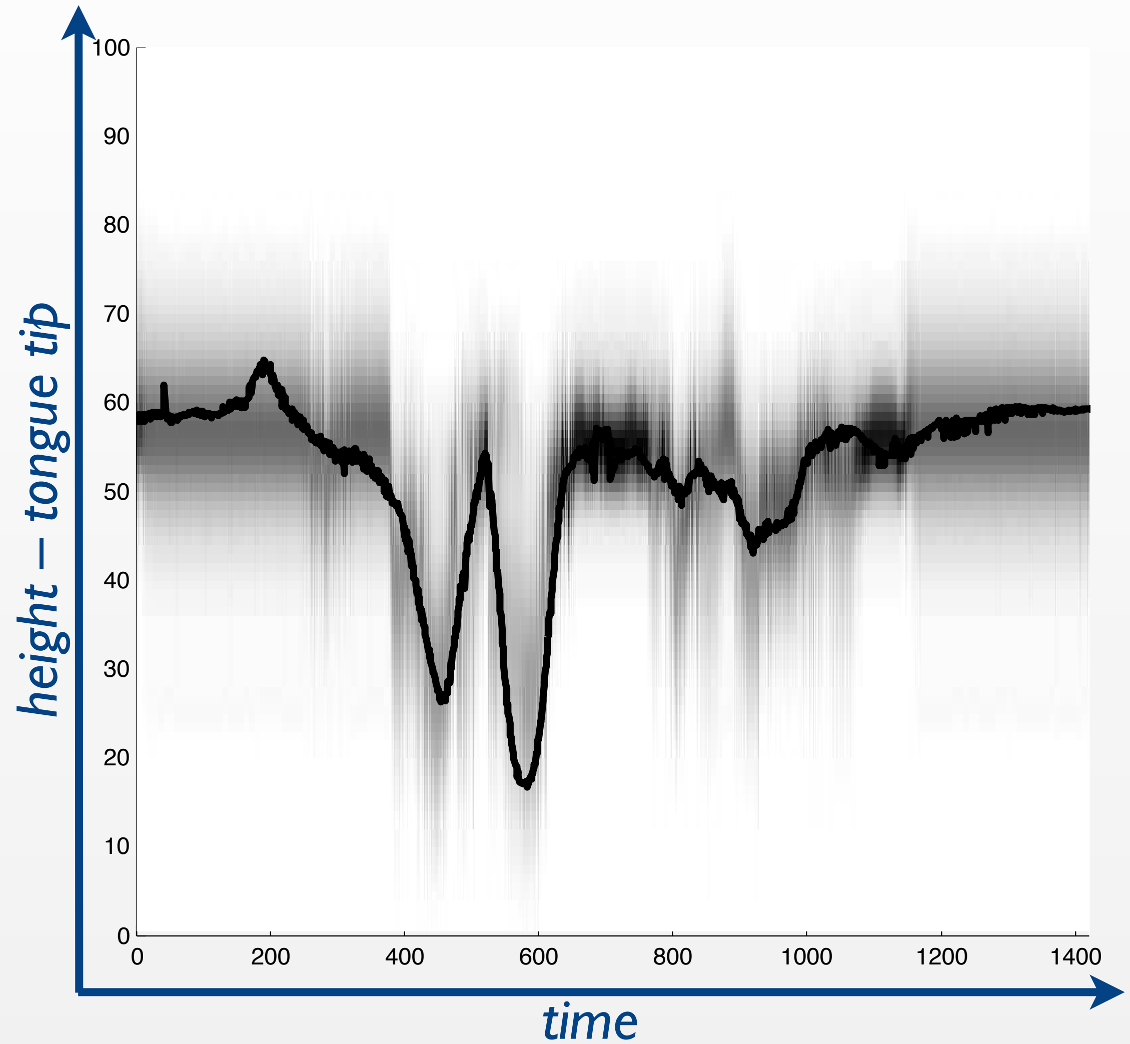
UNIVERSITY of TORONTO

Success of articulatory models

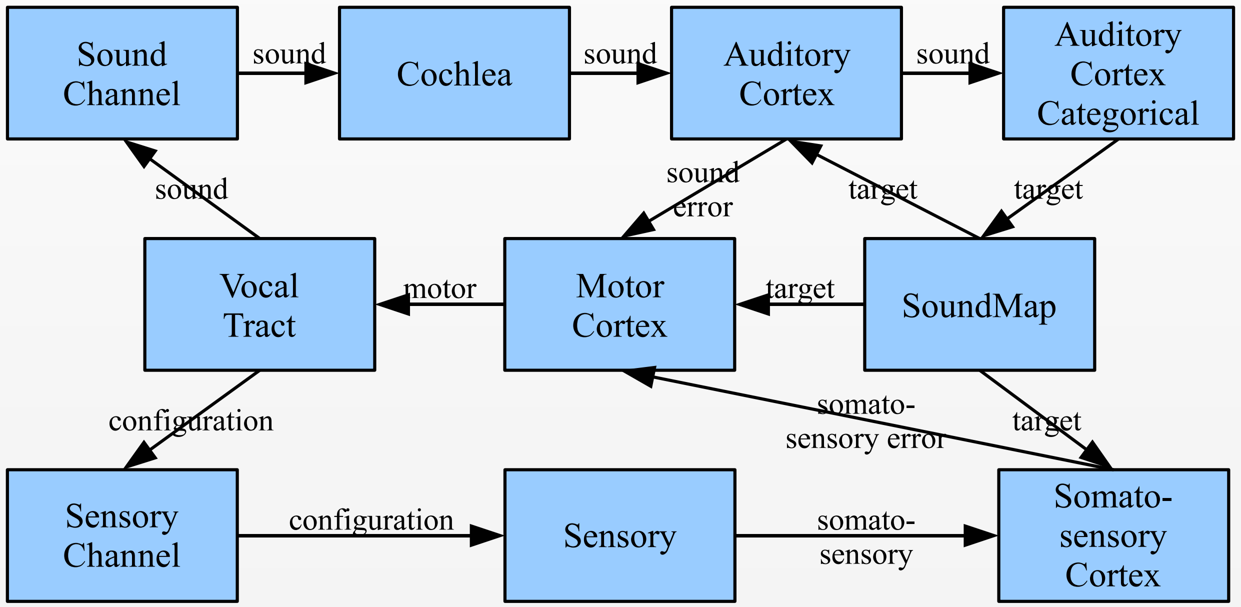
Model	Data	% Accuracy (phonemes)	
		Dysarthric	Control
Baseline (Hidden Markov Model)	-	14.1	72.8
Neural network	Phonological	14.4	72.6
	EMA	16.1	72.7
Dynamic Bayesian network	Phonological	15.0	73.3
	EMA	16.5	73.8
Conditional random field	Phonological	16.5	73.4
	EMA	16.8	73.5

Statistically significant improvement in phoneme recognition.

Current work

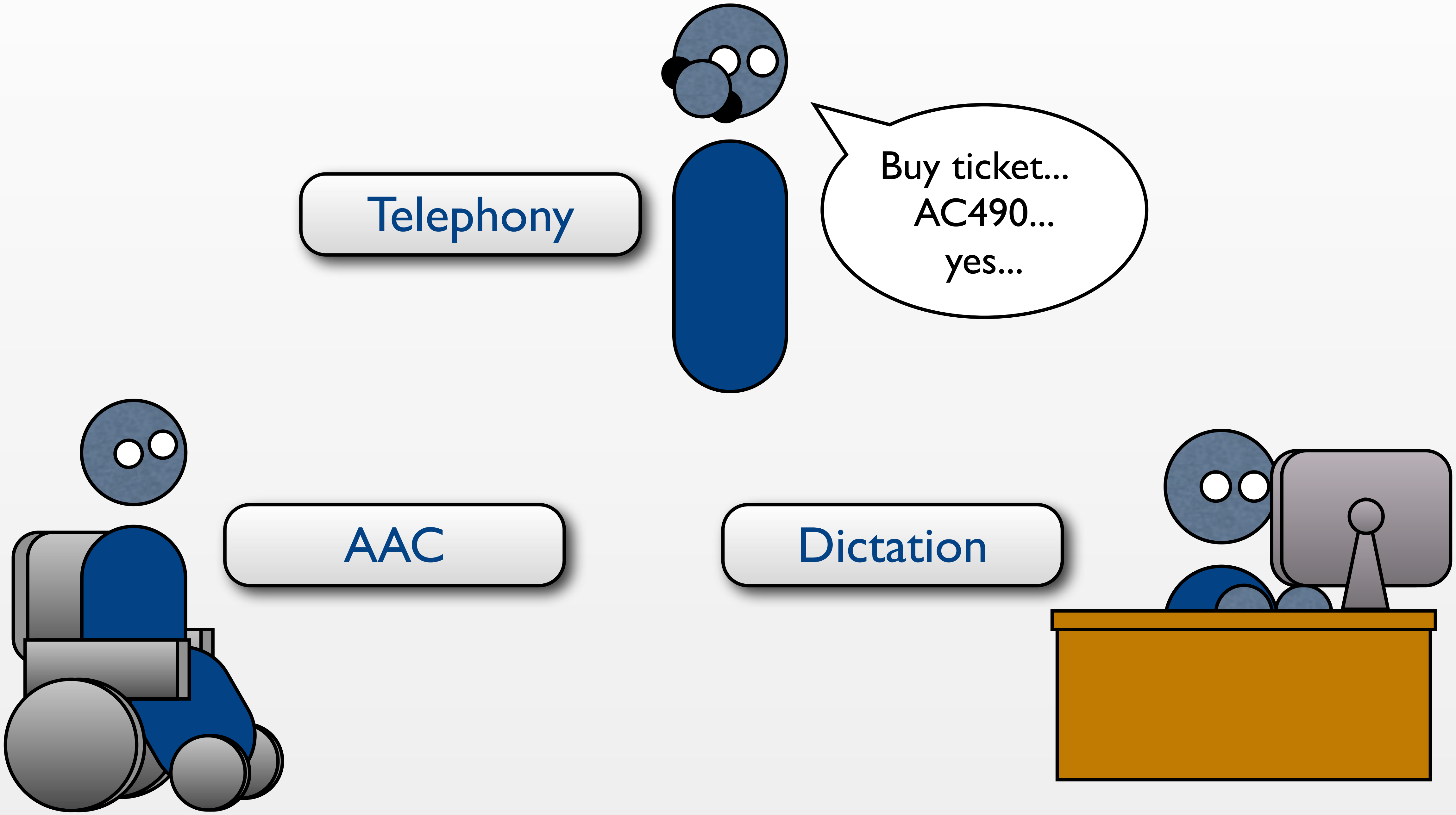


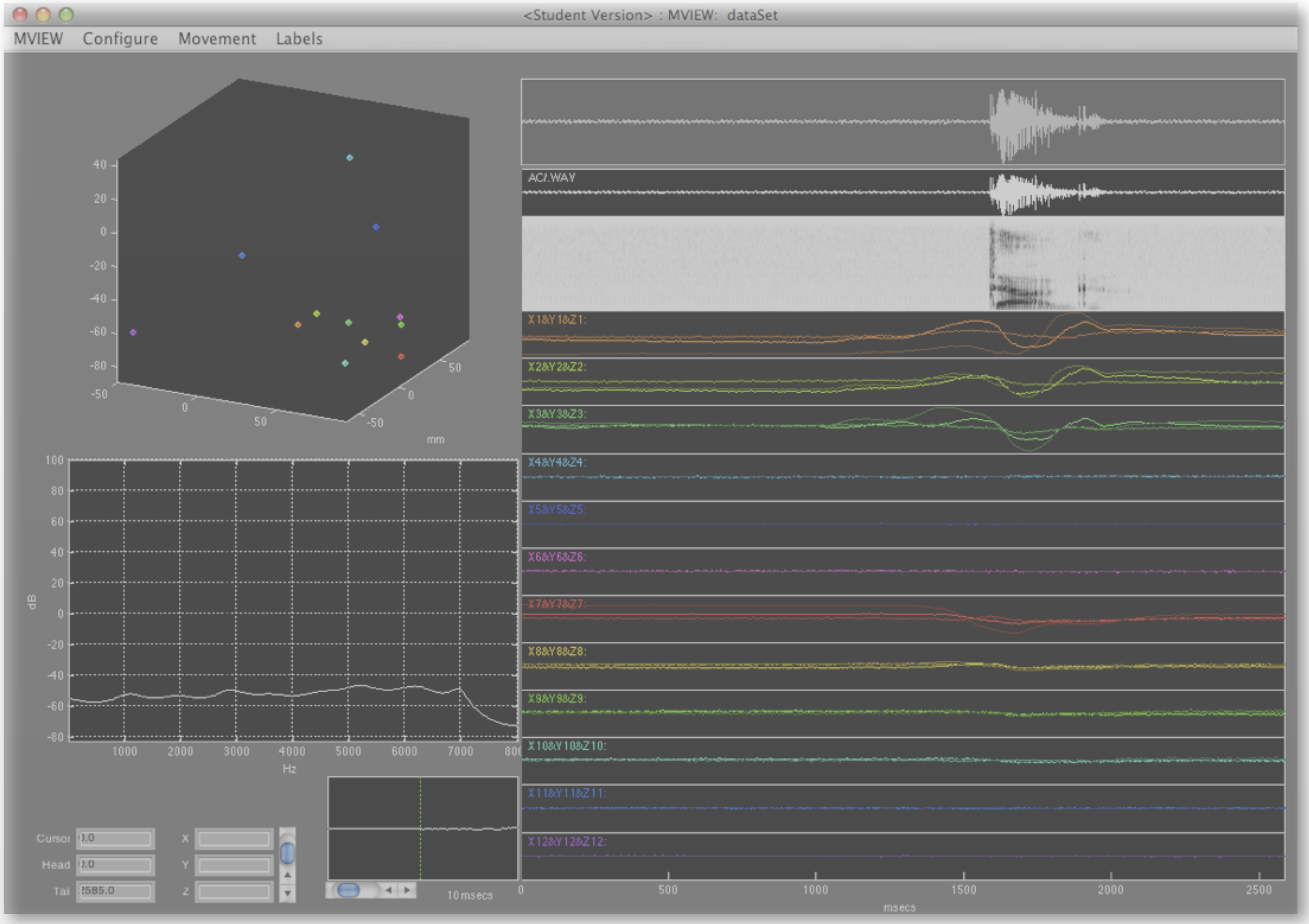
Estimate articulator positions given only acoustics.



More sophisticated models of speech production and recognition.

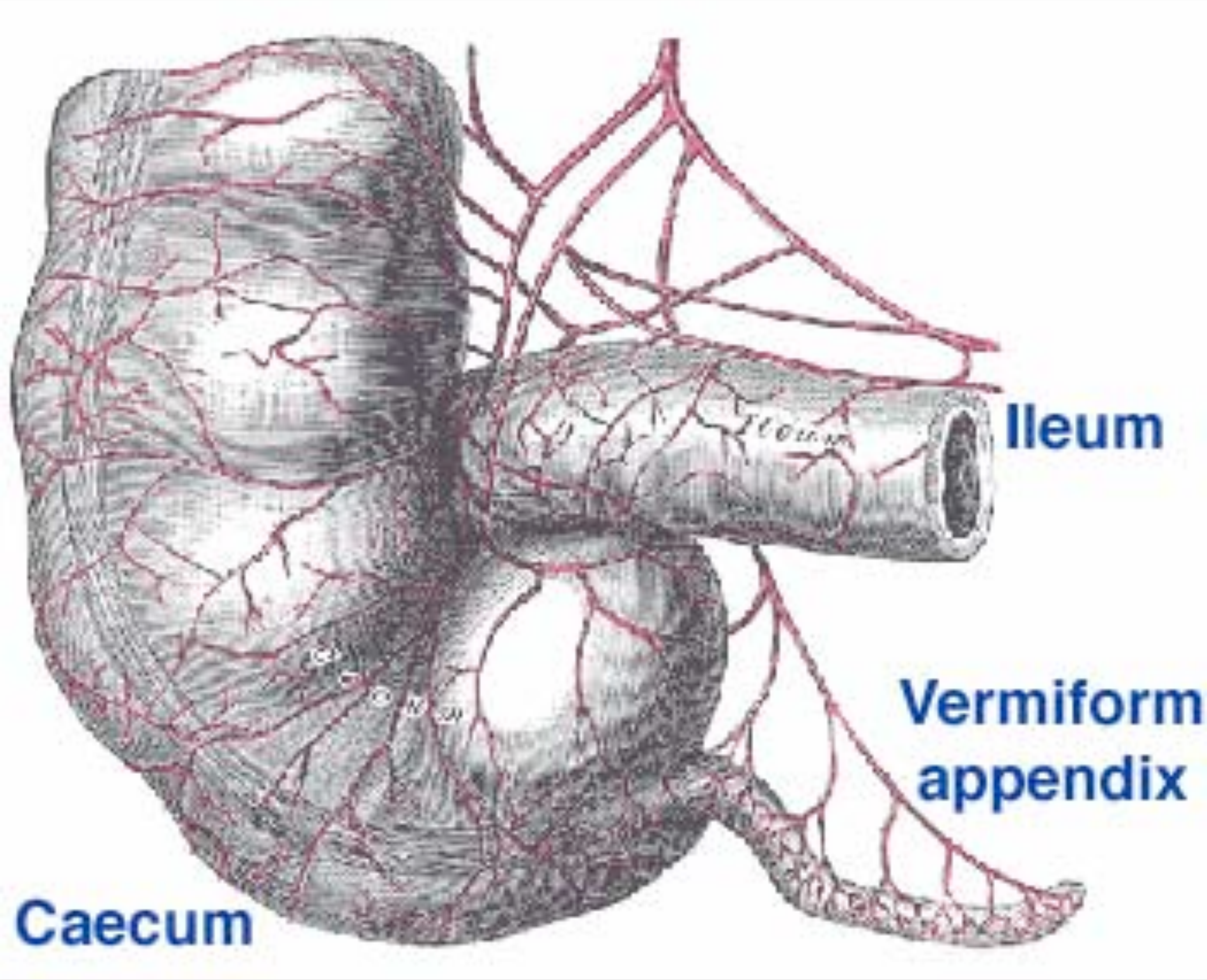
Future work




<Student Version> : MVIEW: dataSet
MVIEW
Configure
Movement
Labels
ACI.WAV
Hz
dB
mm
msecs
Cursor
Head
Tail
10 msecs

Fin

Appendices

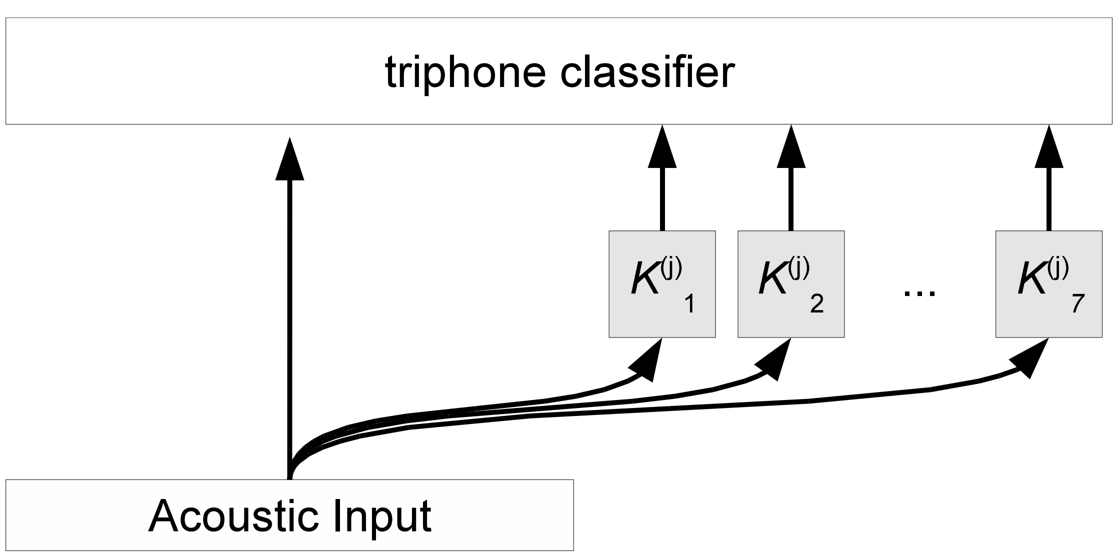


Nosology of the dysarthrias

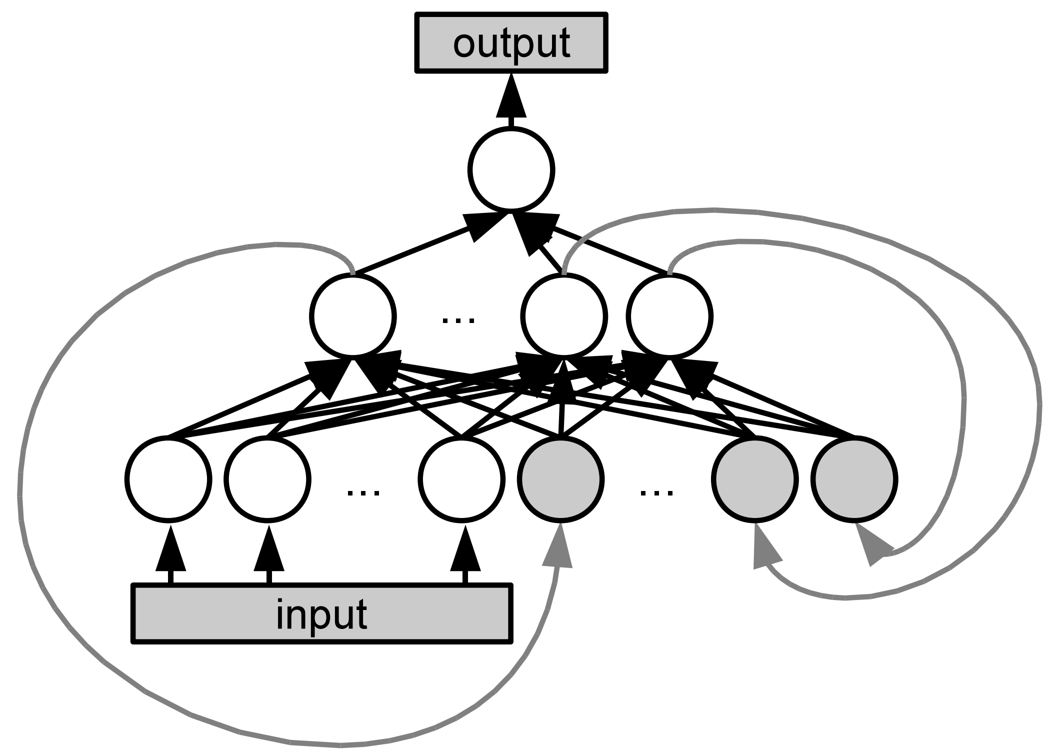
Type	Primary lesion site	Perceptual
Ataxic	cerebellum	Articulatory inaccuracy, prosodic excess, phonatory insufficiency
Flaccid	lower motor neurons	Phonatory incompetence, resonatory incompetence
Spastic	upper motor neurons	Prosodic excess, articulatory-resonatory incompetence
Spastic-Flaccid	upper/lower motor neurons	Prosodic excess, articulatory-resonatory incompetence , phonatory stenosis
Hyperkinetic (chorea)	Basal ganglia (especially putamen or caudate)	articulatory-resonatory incompetence, phonatory stenosis
Hyperkinetic (distonia)	ibid	articulatory inaccuracy, prosodic excess, phonatory stenosis
Hypokinetic	Basal ganglia (substantia nigra)	Prosodic insufficiency, phonatory incompetence

Discriminative classification

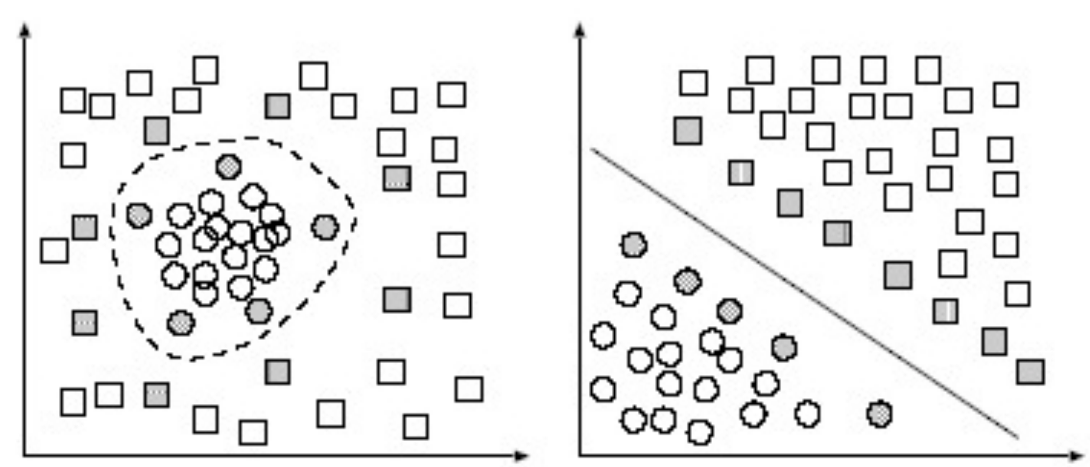
- Neural networks
  - Multi-layer perceptron.
  - Recurrent Elman network.
- Support vector machines
  - Radial-basis function kernel.
  - Dynamic-time warped kernel.
- We discriminatively identify PFs, and combine these classifiers to identify triphones.



NN



SVM



(a) Radial Basis Function

(b) RBF mapping

Phonological features

Feature (PF)	Values
Manner	approximant, fricative, nasal, retroflex, silence, stop, vowel
Place	alveolar, bilabial, dental, labiodental, silence, velar, nil
High/Low	high, mid, low, silence, nil
Voice	voiced, unvoiced
Front/Back	front, central, back, nil
Round	round, non-round, nil
Static	static, dynamic