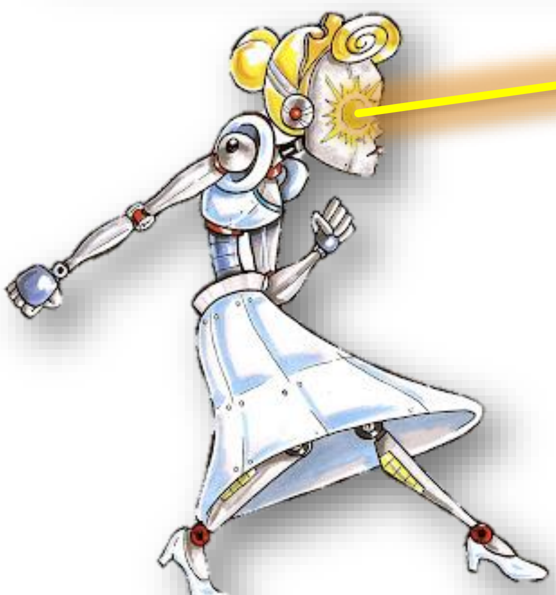
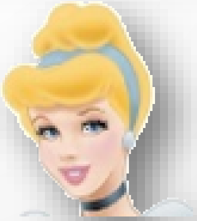
Tongues, brains, Cinderella, and robots
Different ways to handle atypical speech

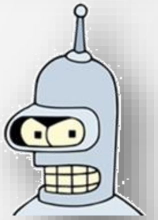


Frank Rudzicz
Scientist, Toronto Rehabilitation Institute
Assistant professor, Department of Computer Science
University of Toronto

18 November 2014, Cambridge MA

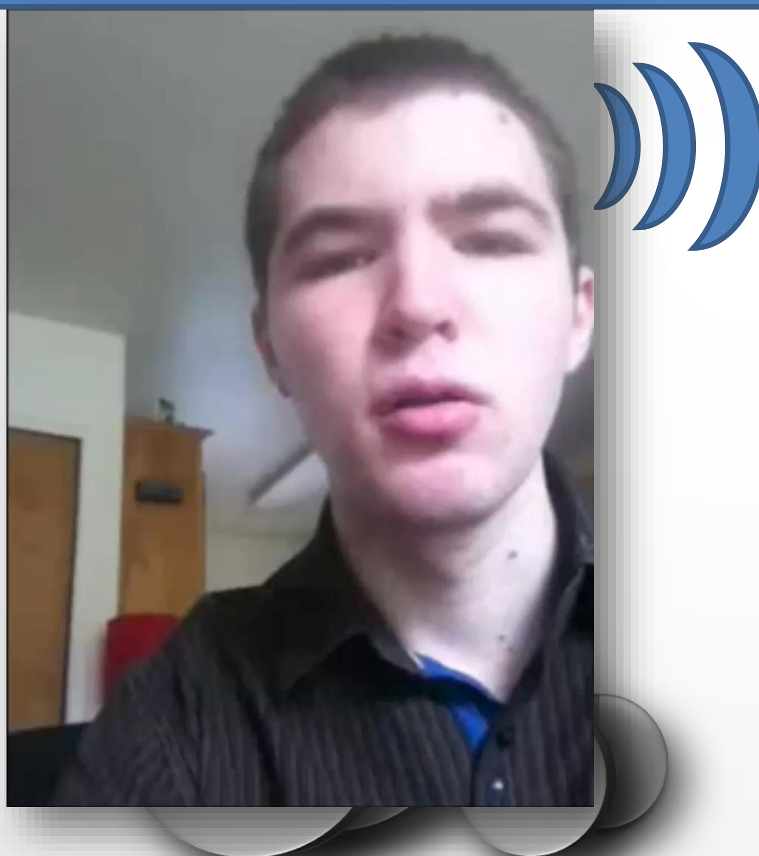
This talk

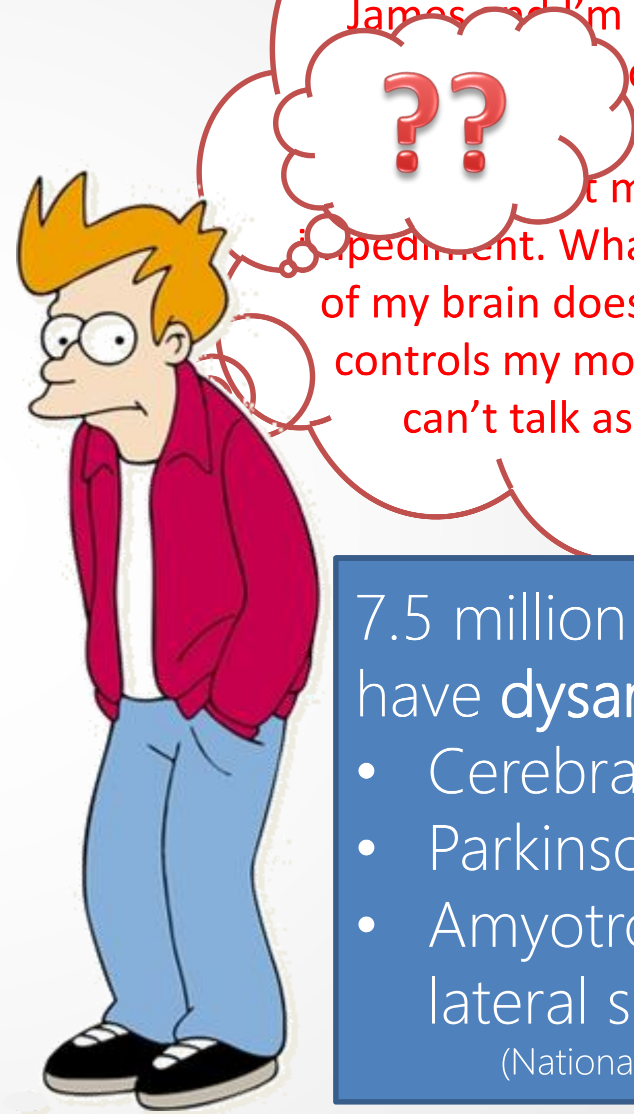




Dysarthria

Neuro-motor articulatory disorders resulting in **unintelligible** speech.





Hey everybody! My name's James ... I'm here to do a ... ch video for ... briefly gonna ... t my speech ... pediment. What it is, is a part of my brain doesn't work that controls my mouth and I um can't talk as perfectly

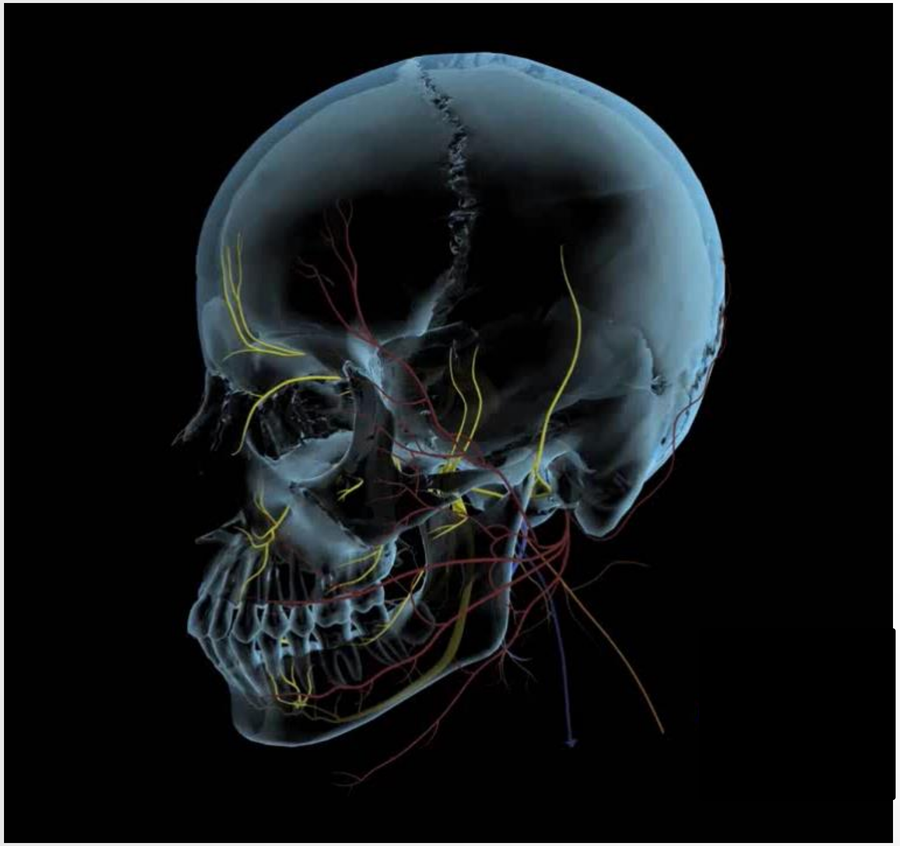
7.5 million Americans have **dysarthria**

- Cerebral palsy,
- Parkinson's,
- Amyotrophic lateral sclerosis

(National Institute of Health)

Nosology of dysarthria

- **Types** of dysarthria are related to **specific sites** in the subcortical nervous system.



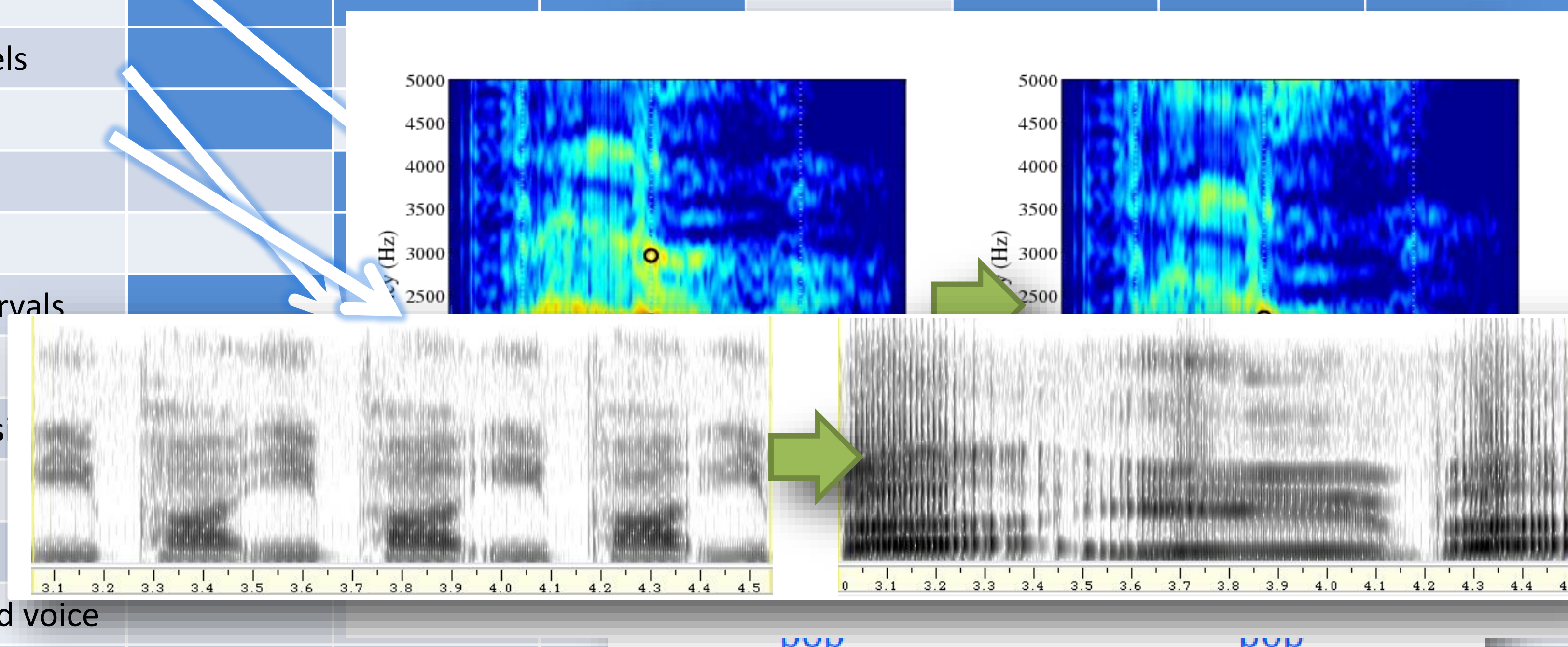
Type	Primary lesion site
Ataxic	**Cerebellum** or its outflow pathways
Flaccid	**Lower motor neuron** (≥1 cranial nerves)
Hypo-kinetic	**Basal ganglia** (esp. substantia nigra)
Hyper-kinetic	**Basal ganglia** (esp. putamen or caudate)
Spastic	**Upper motor neuron**
Spastic-flaccid	Both **upper** and **lower motor neurons**

(After Darley *et al.*, 1969)

SPOClab signal processing and oral communication
UHN Toronto Rehabilitation Institute
University of Toronto

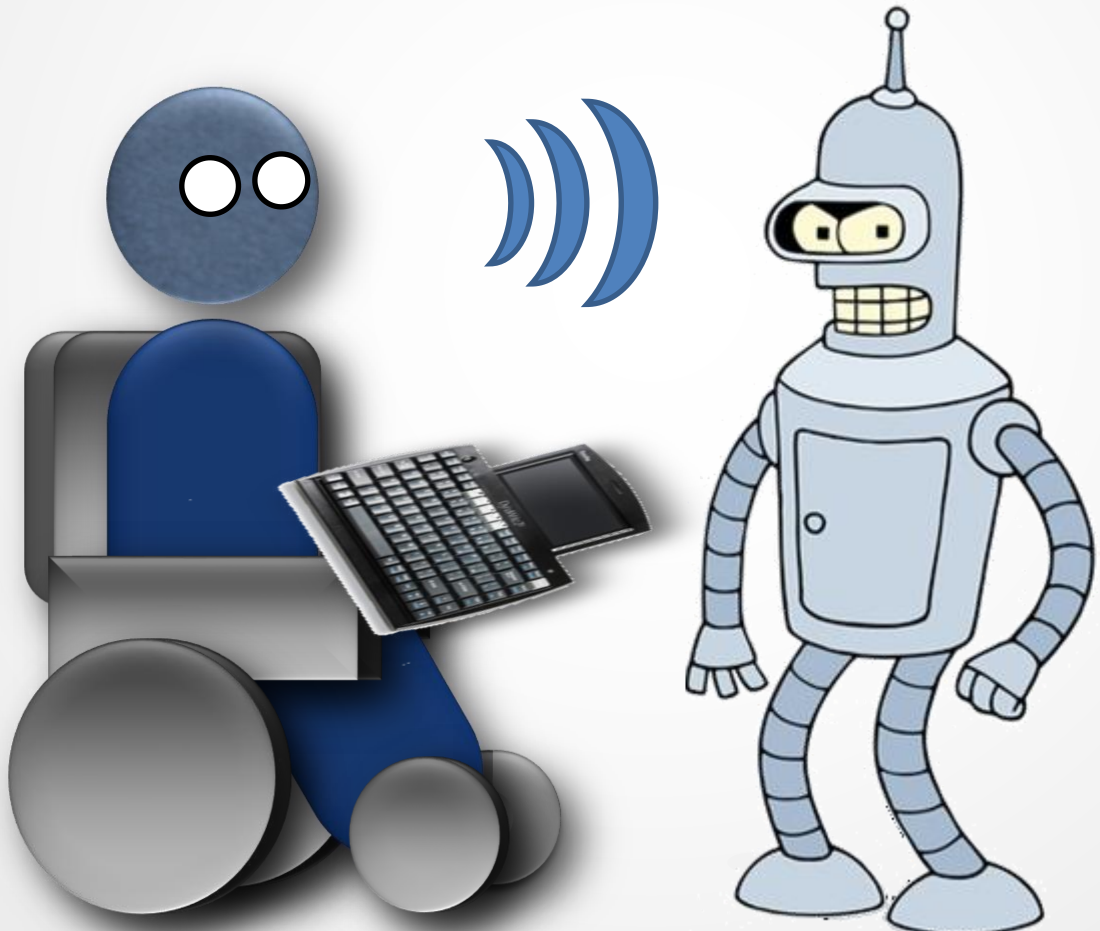
Characteristics of dysarthria

	Ataxic	Flaccid	Hypo-kinetic	Hyper-kinetic, chorea	Hyper-kinetic, dystonia	Spastic	Spastic-flaccid (ALS)
Monopitch							
Harshness							
Imprecise consonants							
Mono-loud							
Distorted vowels							
Slow rate							
Short phrases							
Hypernasal							
Prolonged intervals							
Low pitch							
Inappropriate s							
Variable rate							
Breathy voice							
Strain-strangled voice							
…							



Dysarthria

The **broader** neuro-motor deficits associated with dysarthria can make **traditional** human-computer interaction difficult.

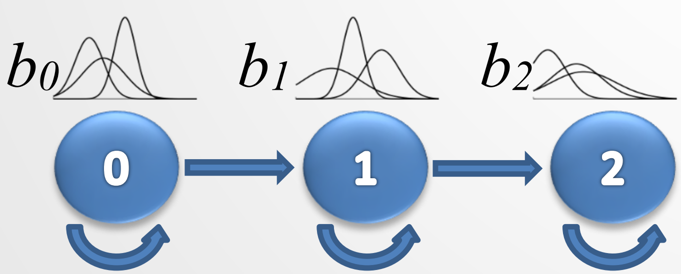


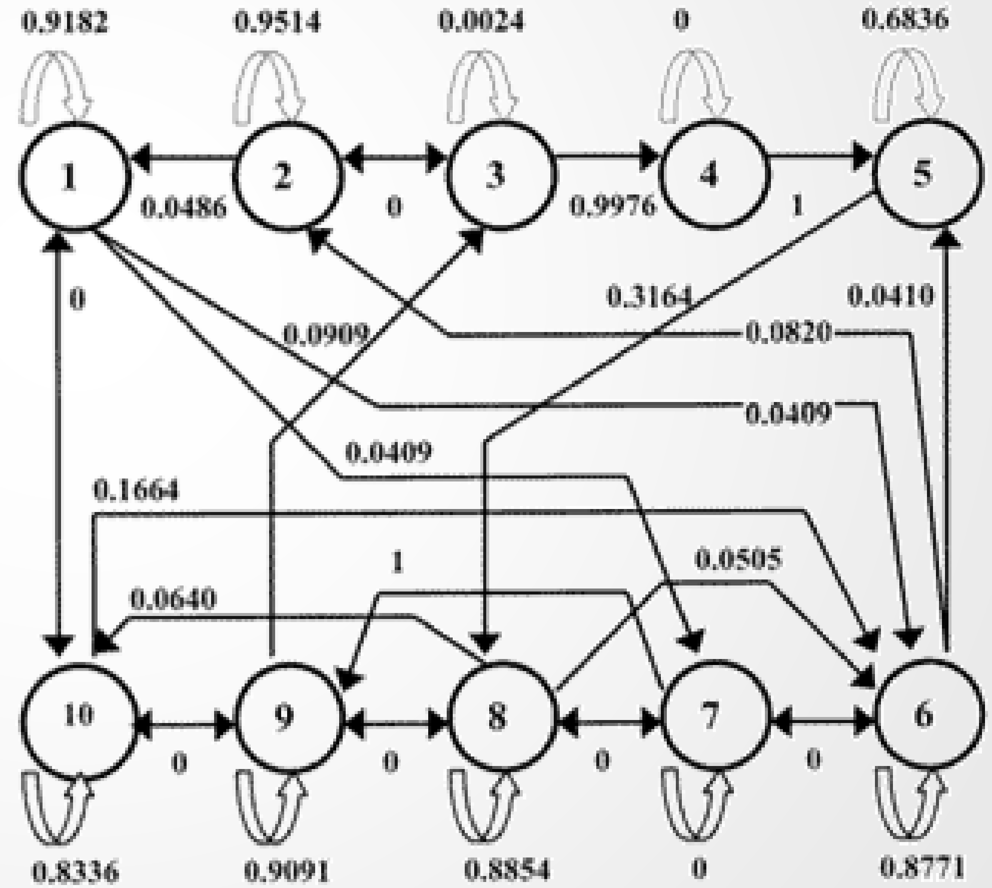
Can we use ASR for dysarthria?

Accounting for aspects of dysarthria

- **Ergodic** HMMs can be **robust** against recurring **pauses**, and **non-speech** events.

- Polur and Miller (2005) **replaced GMM** densities **with neural networks** (after Jayaram and Abdelhamied, 1995), further **increasing accuracy**.





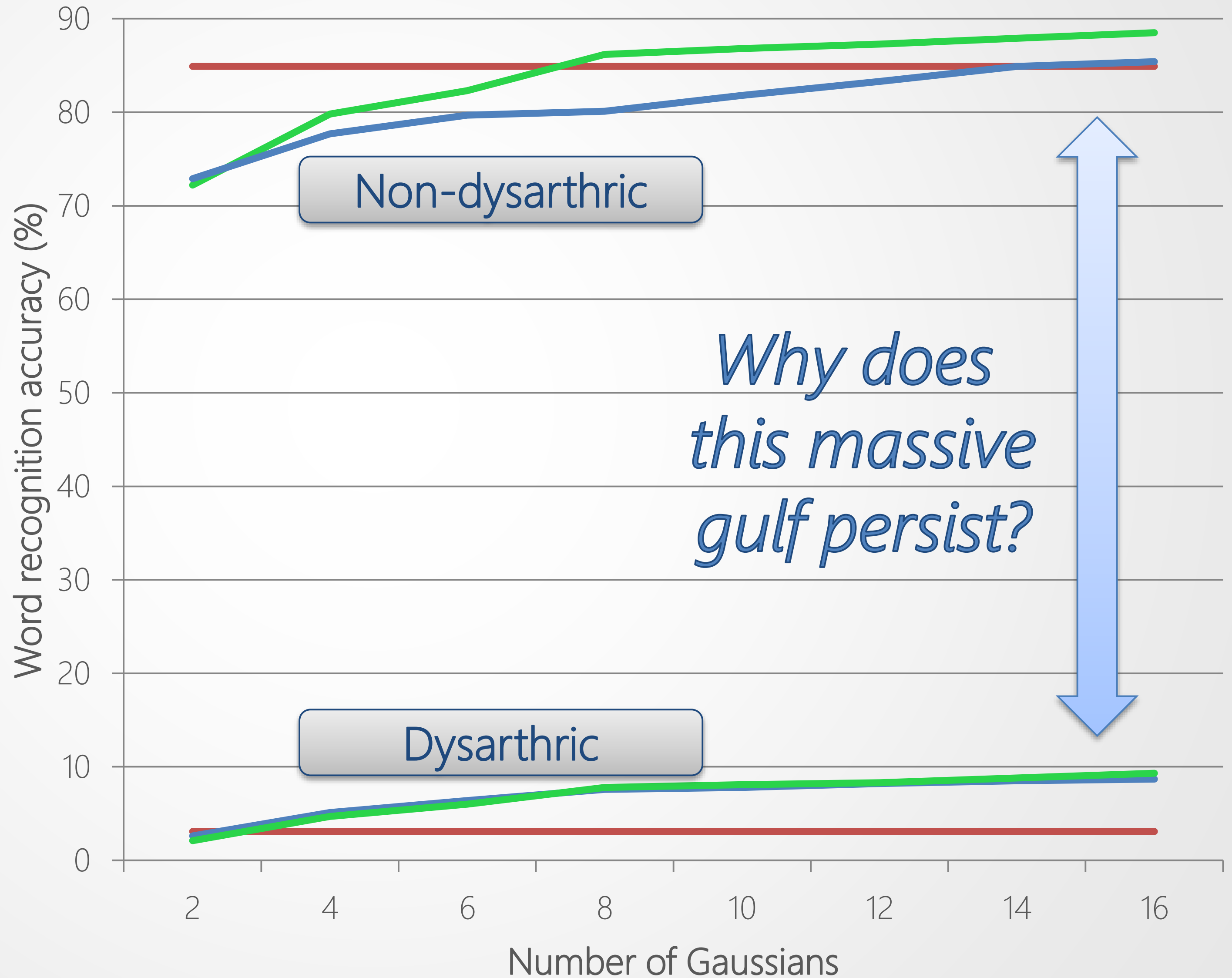
(From Polur and Miller., 2005)

Adjusting to the individual speaker

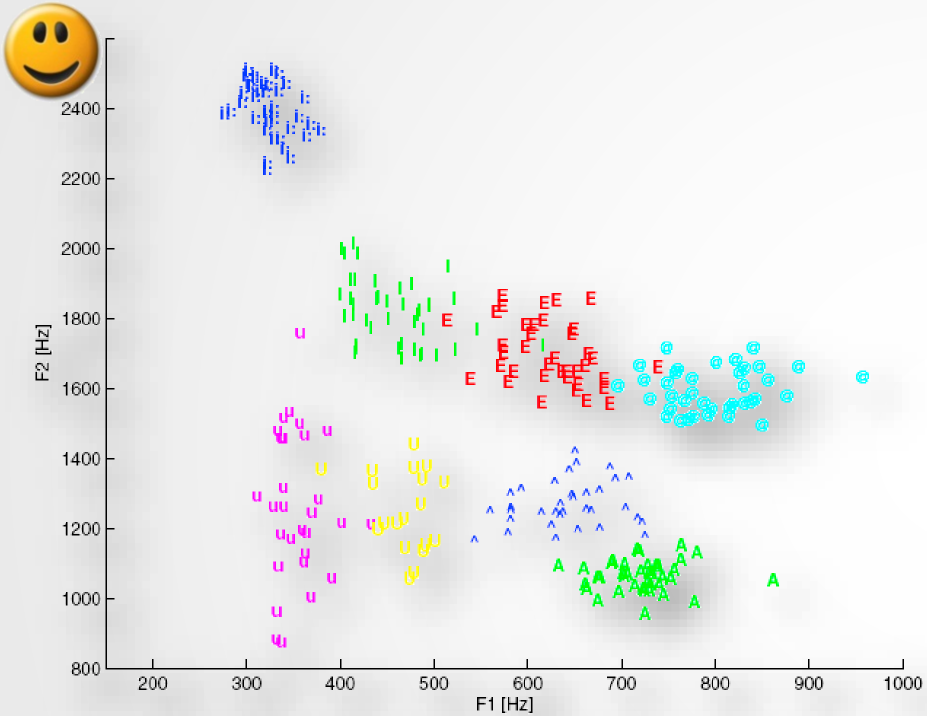
84.9%

Traditional ASR
Speaker-dependent
Speaker-retrained

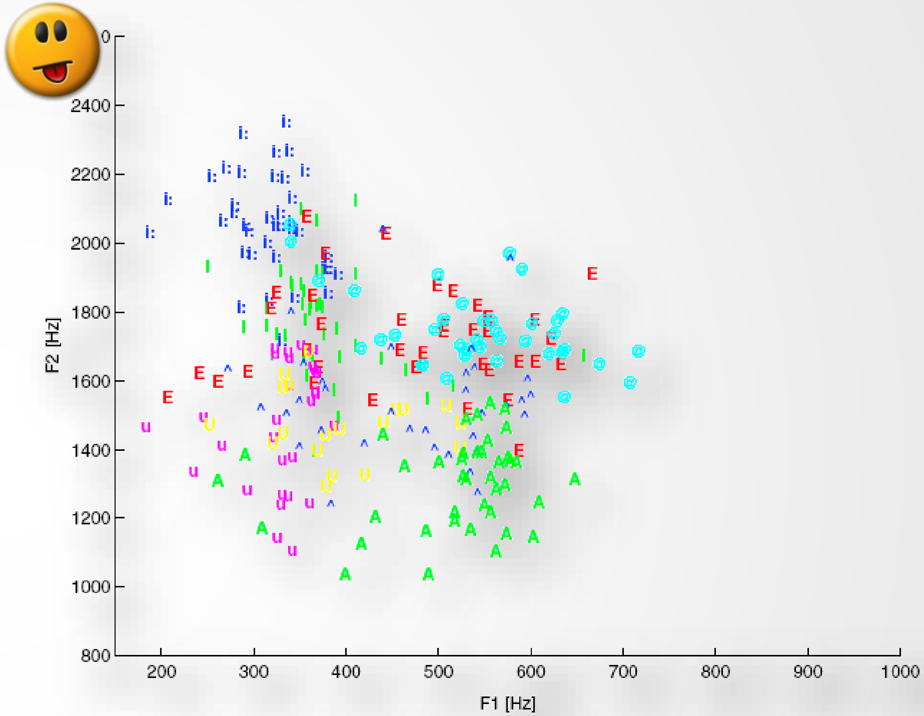
3.1%



Acoustic ambiguity



Non-dysarthric

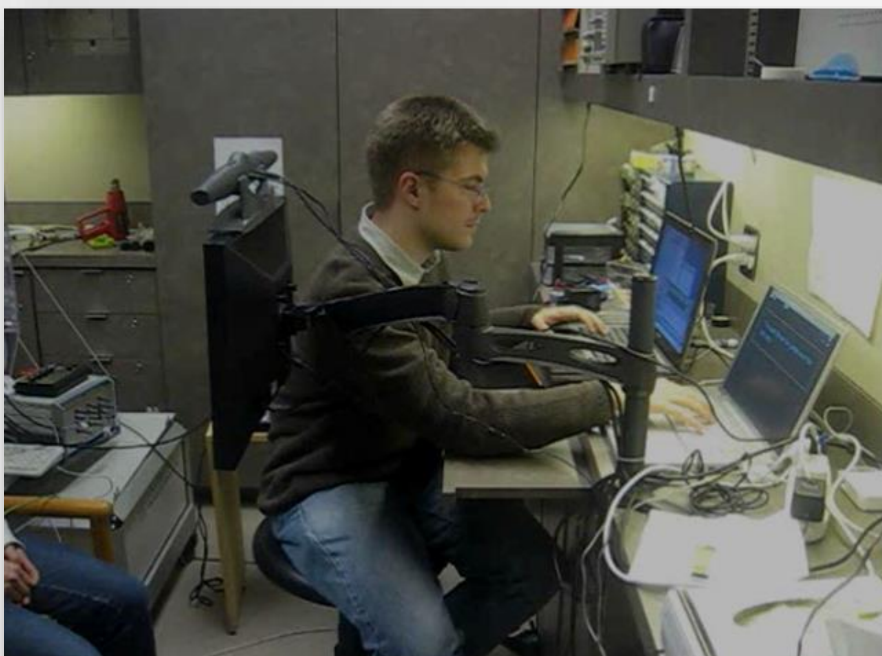


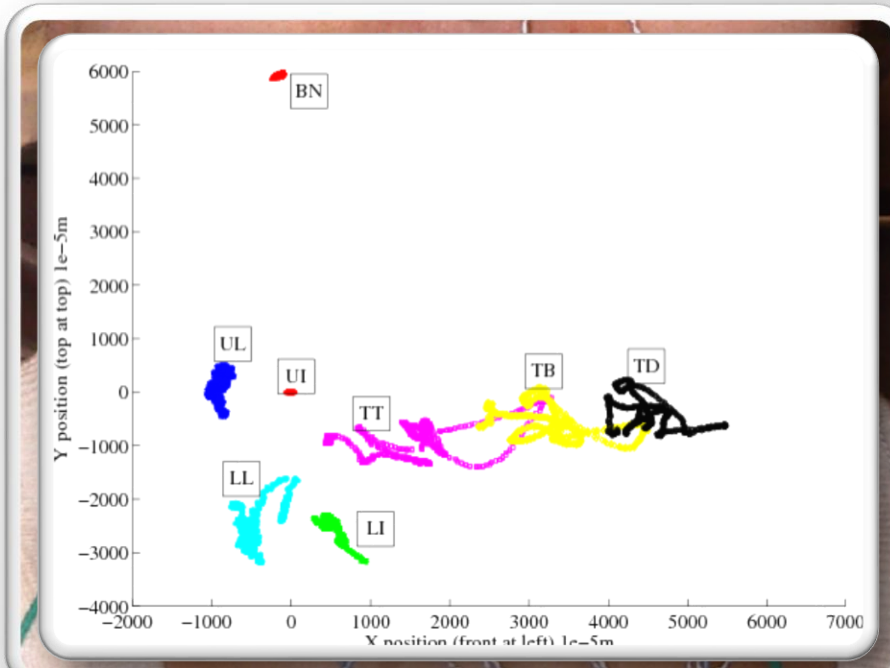
Dysarthric

This **acoustic** behaviour is indicative of underlying **articulatory** behaviour.

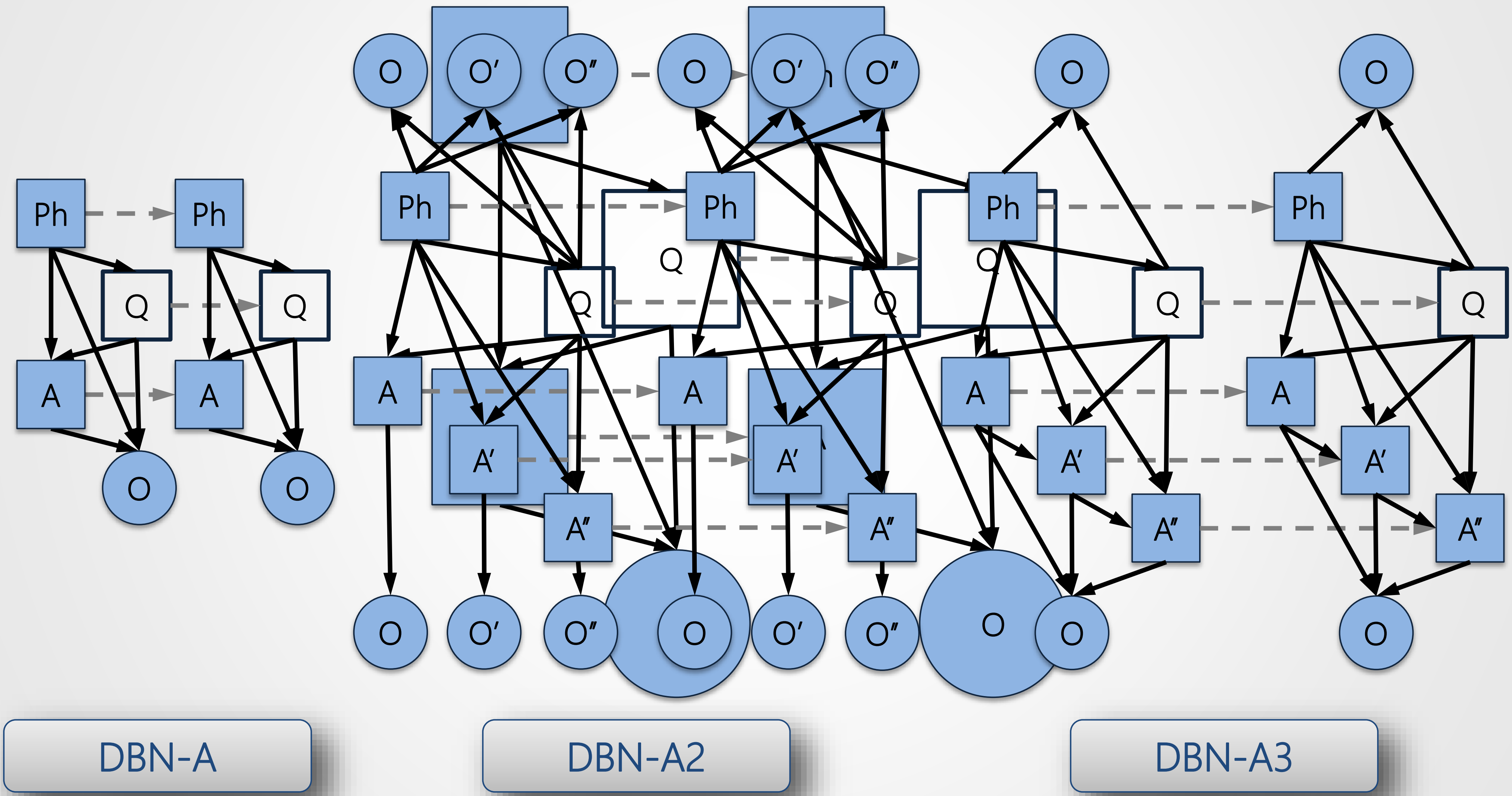
The TORGO database

- TORGO was built to train augmented ASR systems.
  - 9 subjects **with cerebral palsy (1 with ALS)**, **9 matched controls**.
  - Each reads 500—1000 prompts over **3 hours** that cover **phonemes** and **articulatory contrasts** (e.g., *meat* vs. *beat*).
  - **Electromagnetic articulography** (and video) track points to <1 mm error.

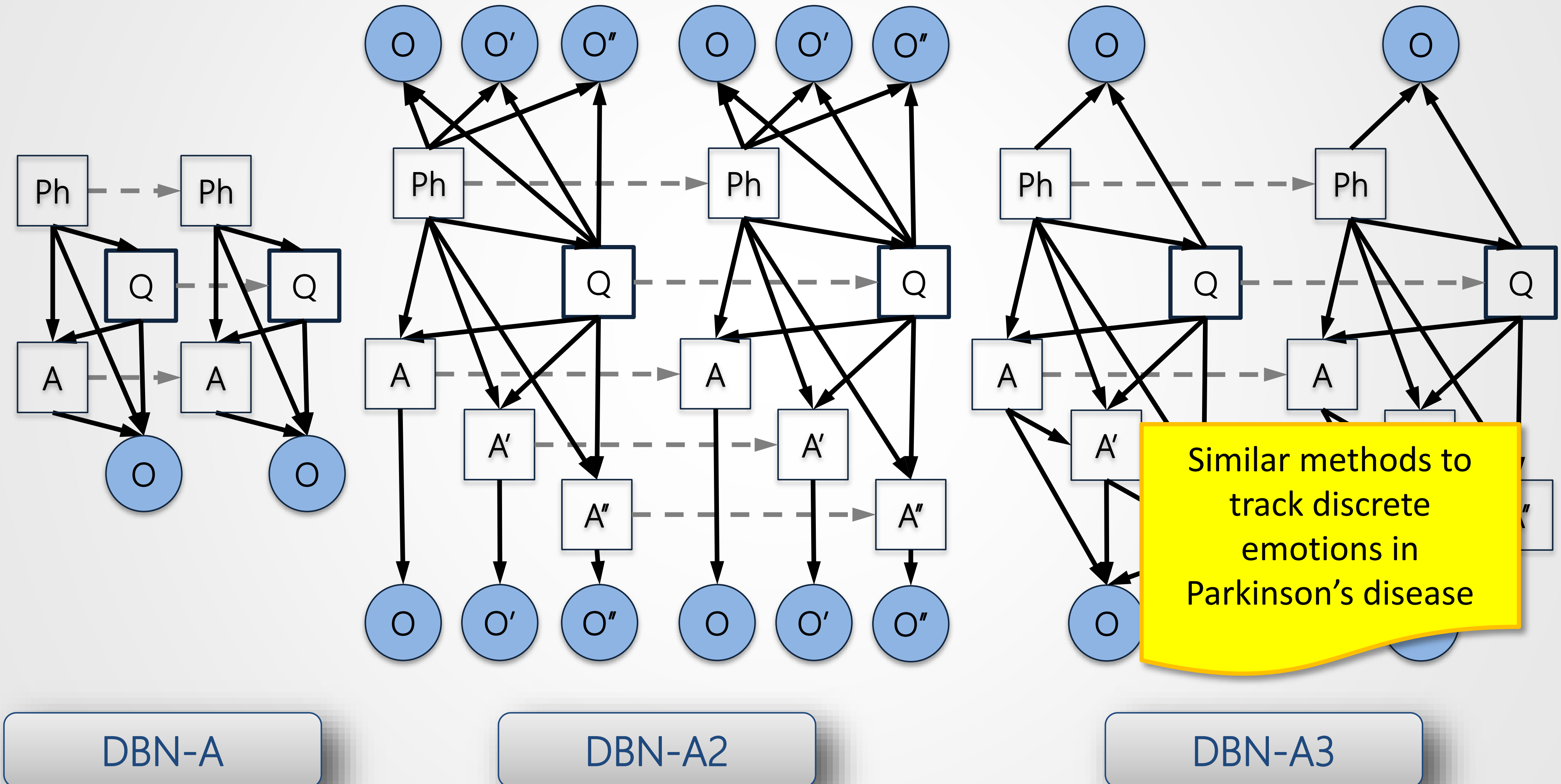




Dynamic Bayes nets with EMA data

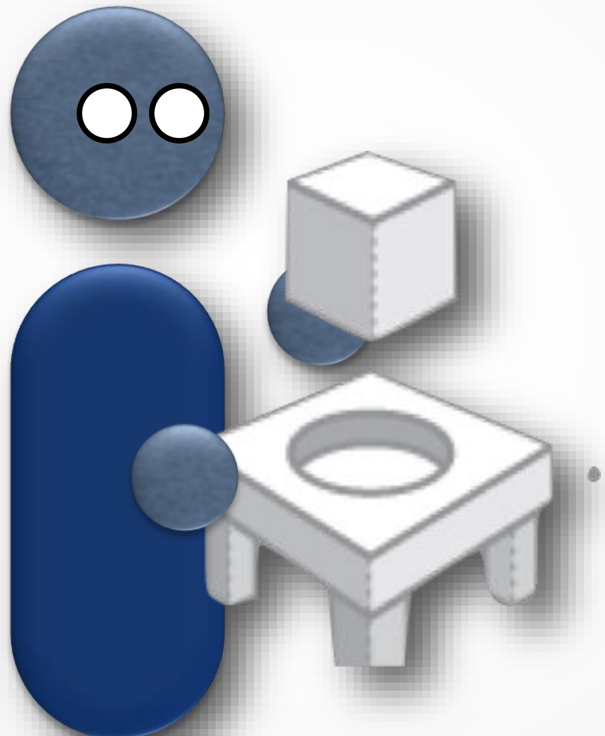


Dynamic Bayes nets with EMA data



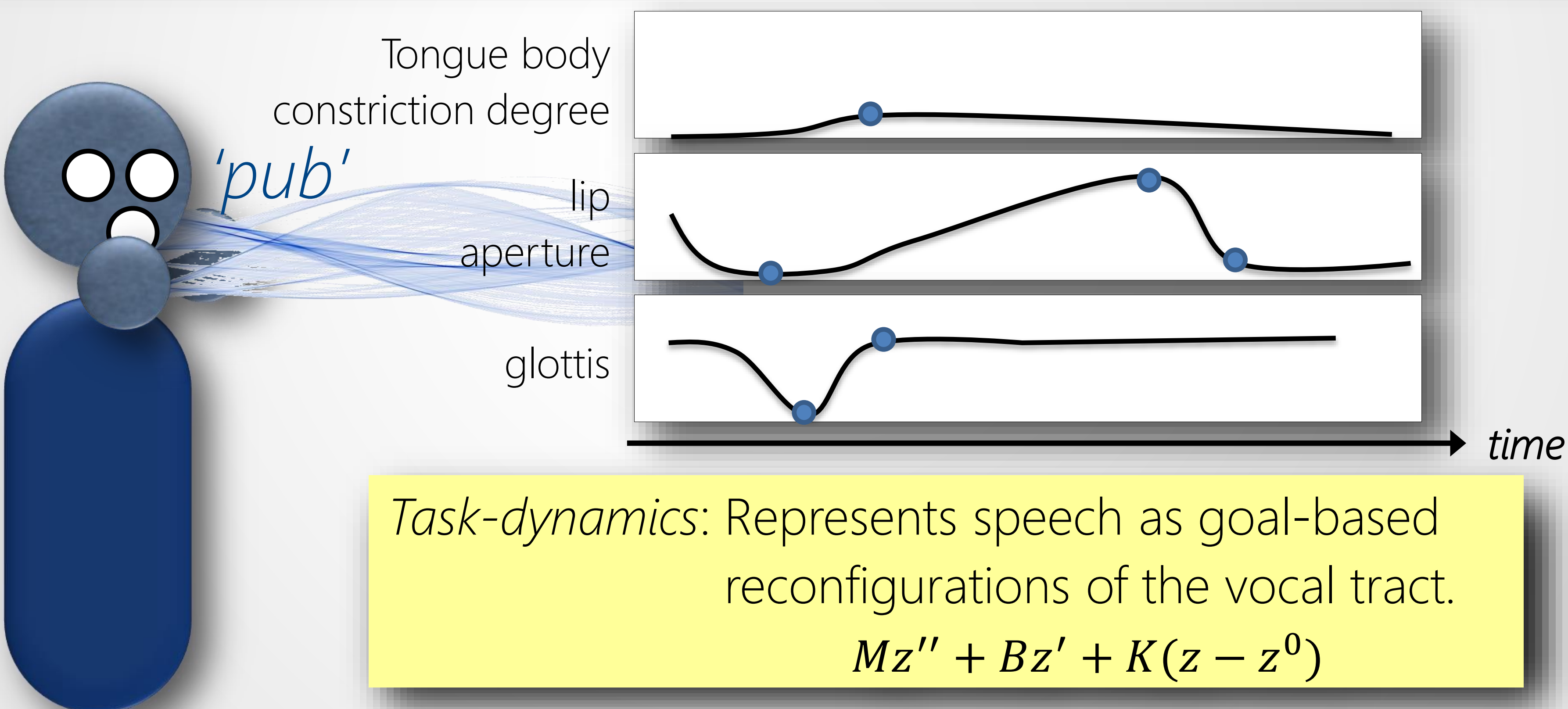
Similar methods to track discrete emotions in Parkinson's disease

Beyond discrete articulation



Dynamic speech gestures

We wish to represent speech in a low-dimensional and informative space that incorporates **goal-based** and **long-term dynamics.**

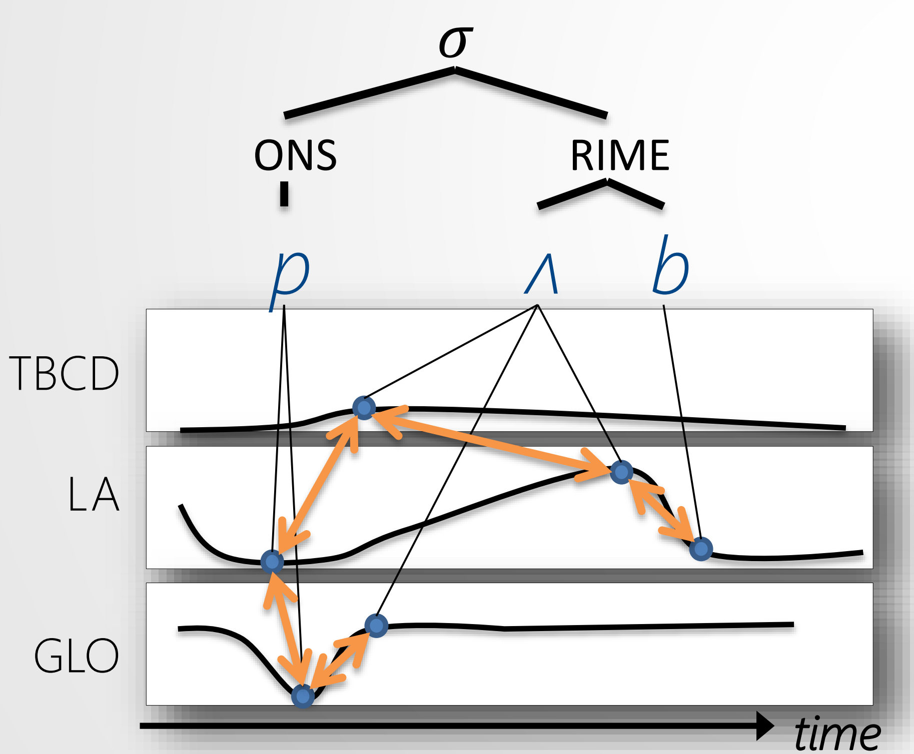


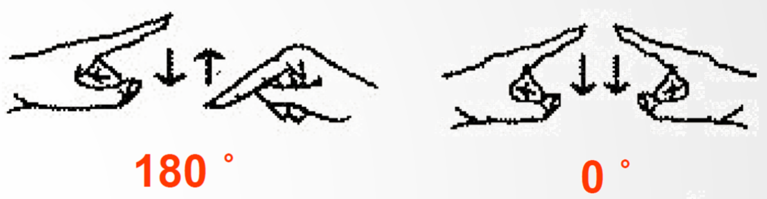
Task-dynamics: Represents speech as goal-based reconfigurations of the vocal tract.

$$Mz'' + Bz' + K(z - z^0)$$

Problem 1: Timing

- In TD, **pairs of goals** are **dynamically coupled** in time.
- Articulators are **phase-locked** (0˚ or 180˚; Goldstein *et al.*, 2005)





- (C)CV pairs stabilize **in-phase**.
- V(C)C pairs stabilize **anti-phase**.
- Kinematic errors occur when competing gestures are **repeated** and tend to stabilize incorrectly.
  - e.g., repeat *koptop* (Nam *et al*, 2010).

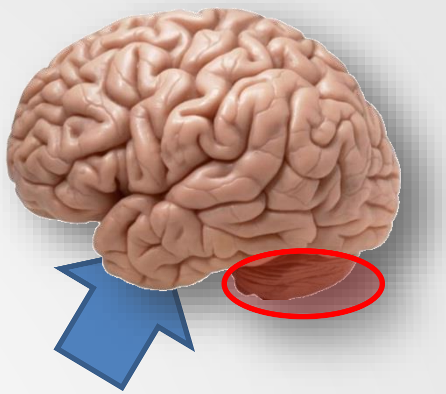
Problem 1.5: Timing/rhythm

- **Rhythm** (the distribution of **emphasis**) is *not* part of TD.
- **Tremor** behaves as oscillations about an equilibrium.
  - There is **evidence** that people with **Parkinson's** coordinate **voluntary** movement with **involuntary** tremors (Kent *et al.*, 2000).
- **Rhythm** in **ataxic** dysarthria formalized by aberrations in a 'scanning index', SI, consisting of syllable lengths S_i,

$$SI = \frac{\prod_{i=1}^{n} S_i}{\left(\frac{\sum_{i=1}^{n} S_i}{n}\right)^n}$$

(Ackermann and Hertrich, 1994))

Problem 2: Feedback

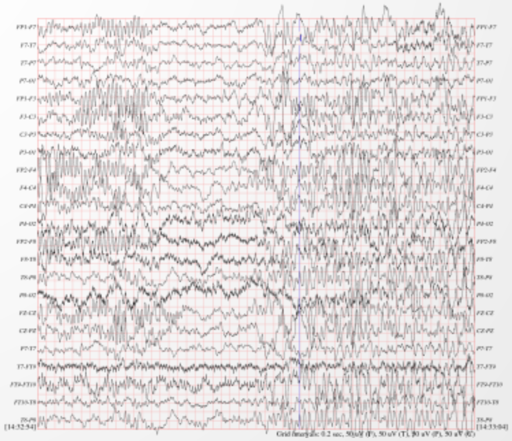


- Dysarthria can affect **sensory** cranial nerves.
- **Parkinson's disease** reduces **temporal** discrimination in **tactile**, **auditory**, and **visual** stimuli.
  - Likely explanation is that **damage** to the basal ganglia **prohibits** the formation of **sensory targets** (Kent *et al.*, 2000).
  - The result is **underestimated** movement.
- **Cerebellar disease** results in **dysmetria** since the **internal model** of the **skeletomuscular system** is **dysfunctional**.
  - The cerebellum is apparently used in the preparation and revision of movements.

Interpreting brain signals

- Many people are **not merely *dysarthric***, but have **locked-in syndrome – they *cannot even move***.

- **HMMs** have been used in **BCI** to classify **EEG** data.
  - What **features** and **sensor locations** are most informative?
  - How to remove **artifacts** from very noisy signals?
  - How to **elicit** imagined words?



Semantics from EEG

- Classify speech stimuli as either **synonymous** or **non-synonymous** with a prior prime in a speech-receptive task using only EEG data with up to **86.84%** accuracy

(Parisotto *et al.*, submitted$_{a}$).

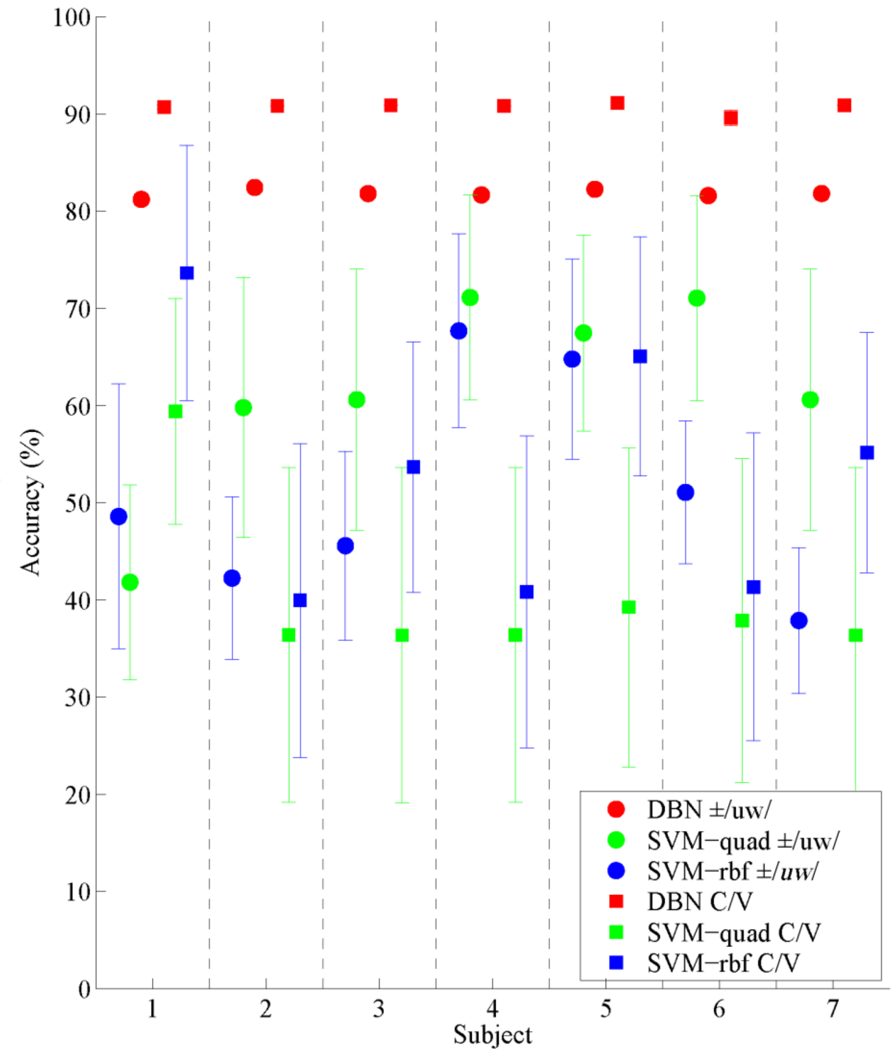
'Semantics' from MEG

- Identify the **language** being received during auditory stimuli in **English** and **Romanian** before and after several weeks of learning words in the latter using MEG, with **>90%** accuracy.

(Parisotto *et al.*, submitted$_b$).

- Significant effects of **semantic word category**, of the subject's ability to play a **musical instrument**, and of the **parietal lobe**.

Phonology from EEG

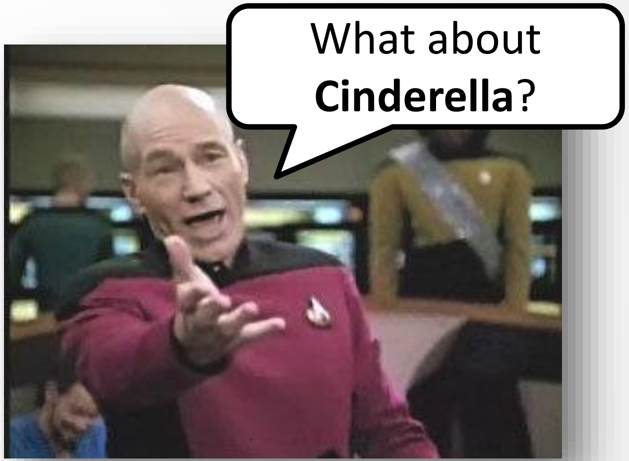


SPOClab
signal processing and
oral communication

UHN Toronto Rehabilitation Institute

UNIVERSITY OF TORONTO


Enough!


What about **Cinderella**?

Further into the brain with aphasia

Broca's aphasia

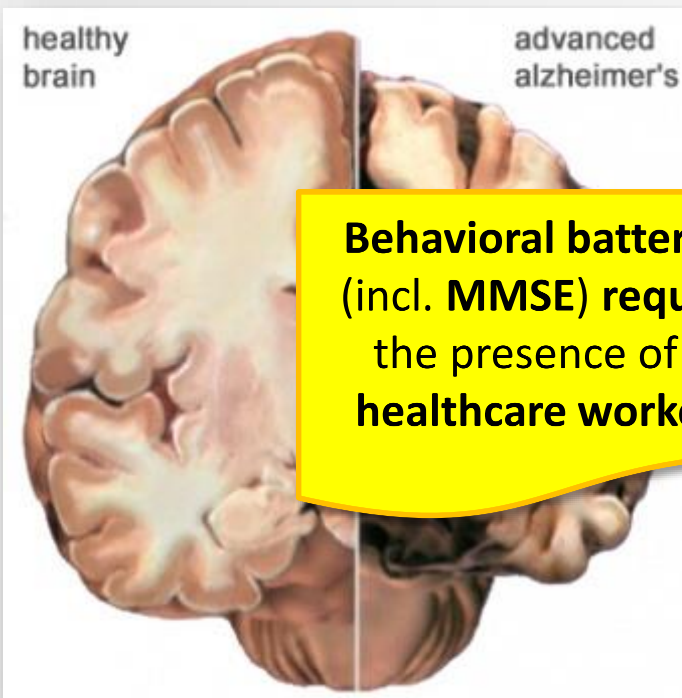
- **Reduced** hierarchical **syntax**.
- **Anomia**.
- **Reduced** "mirroring" between **observation** and **execution** of **gestures** (Rizzolatti & Arbib, 1998).

Wernicke's aphasia

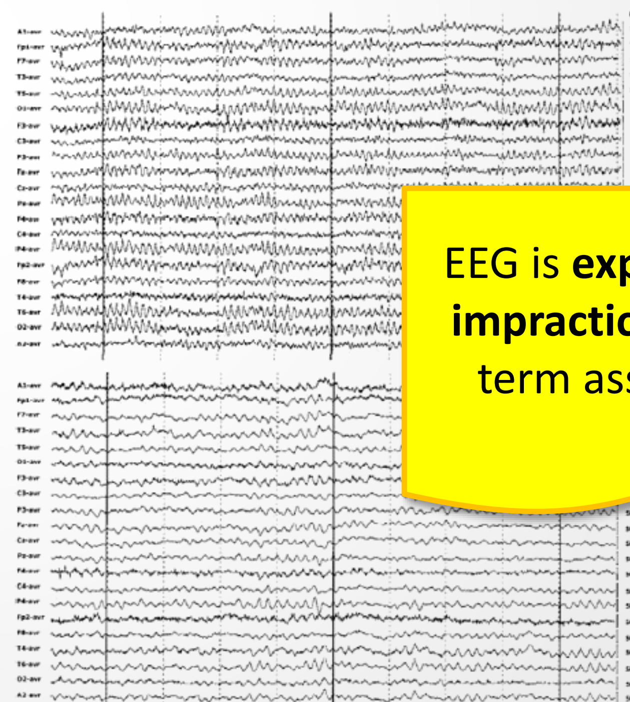
- **Normal** intonation/rhythm.
- **Meaningless** words.
- '**Jumbled**' syntax.
- **Reduced** comprehension.

~~Diagnosis~~ Assessment

- **Alzheimer's disease** (AD) is a progressive neuro-degenerative **dementia** characterized by declines in:
  - Cognitive ability (e.g., **memory**, reasoning),
  - Social ability (e.g., **linguistic** abilities), and
  - Functional capacity (e.g., **executive** power).



Behavioral batteries (incl. **MMSE**) **require** the presence of **a healthcare worker**.

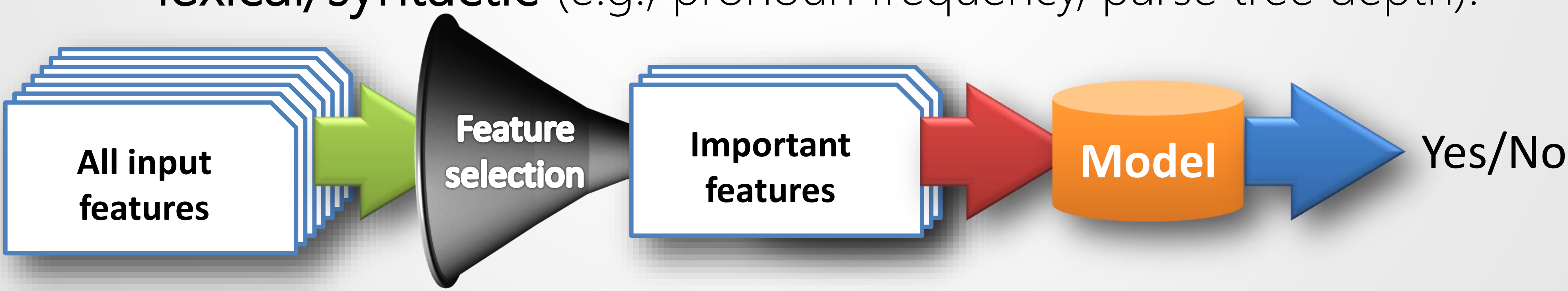


EEG is **expensive** and **impractical** for long-term assessment.

Assessment

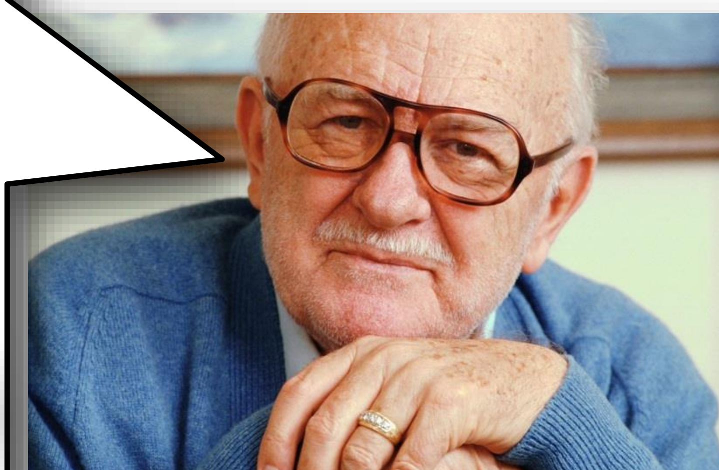
- **Recent work** aims to **identify language disorders**. E.g.,
  - primary progressive aphasia (**PPA**) and its subtypes
    (i.e., semantic dementia (**SD**) and progressive nonfluent aphasia (**PNFA**))
  - **Extended to Parkinson's disease** and **Alzheimer's disease**.

- **Input**: hundreds of features:
  - **acoustic** (e.g., formants, pitch, jitter, shimmer, recurrence) and
  - **lexical/syntactic** (e.g., pronoun frequency, parse tree depth).



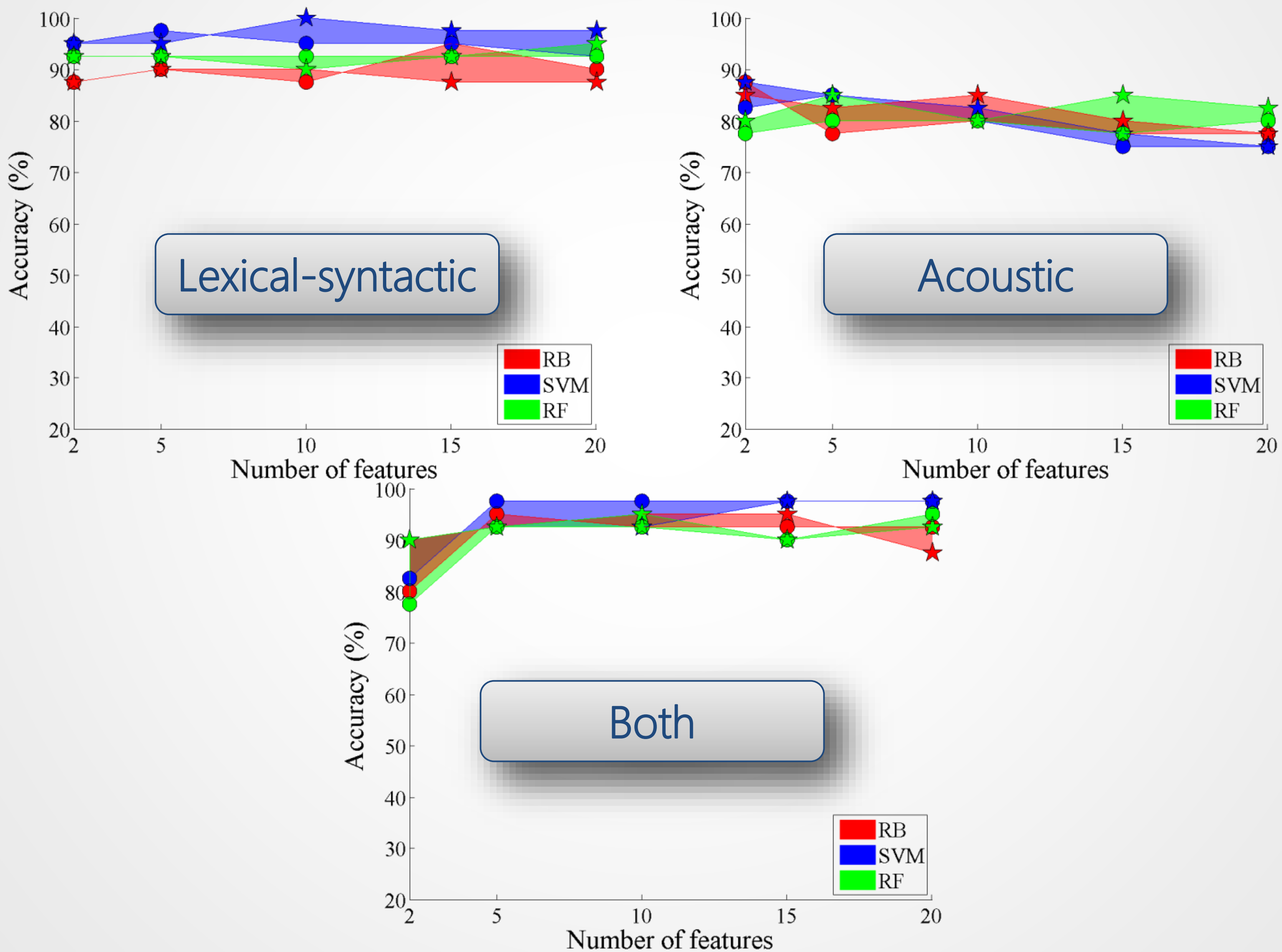
Primary progressive aphasia

- **24 patients** with PPA (14 PNFA, 10 SD) and **16 controls**.
- Narrative recounting of **Cinderella** (after Saffran *et al.* (1989)).
- **Important features**: phonation rate, syntactic complexity, the `familiarity' and frequency of NNs and PRPs, and vocal jitter.



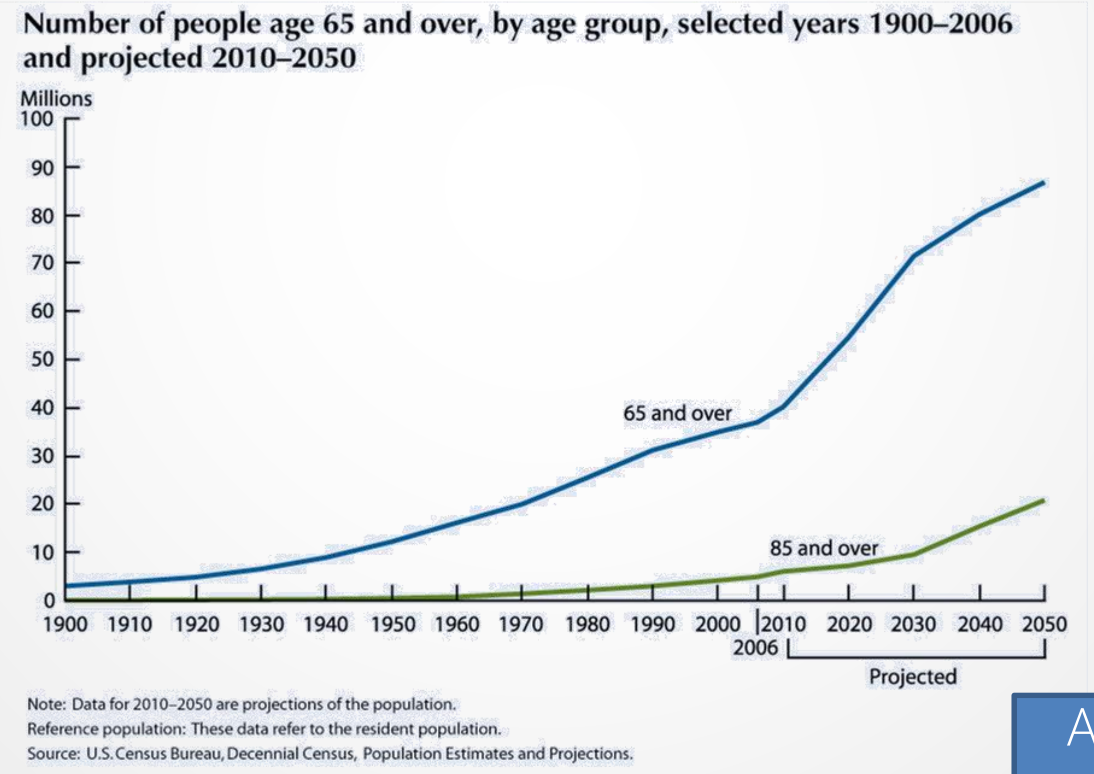
	SD (n=10)	PNFA (n = 14)	Control (n = 16)
Age	65.6 (7.4)	64.9 (10.1)	67.8 (8.2)
Years of edu.	17.5 (6.1)	14.3 (3.6)	16.8 (4.3)
Sex	3 F	6 F	7 F

Identifying PPA



Demographic crisis

- **Caregivers** often assist individuals with AD who live alone, either at **home** or in **long-term care facilities**.
  - >$100B are spent annually in the U.S. on caregiving AD.



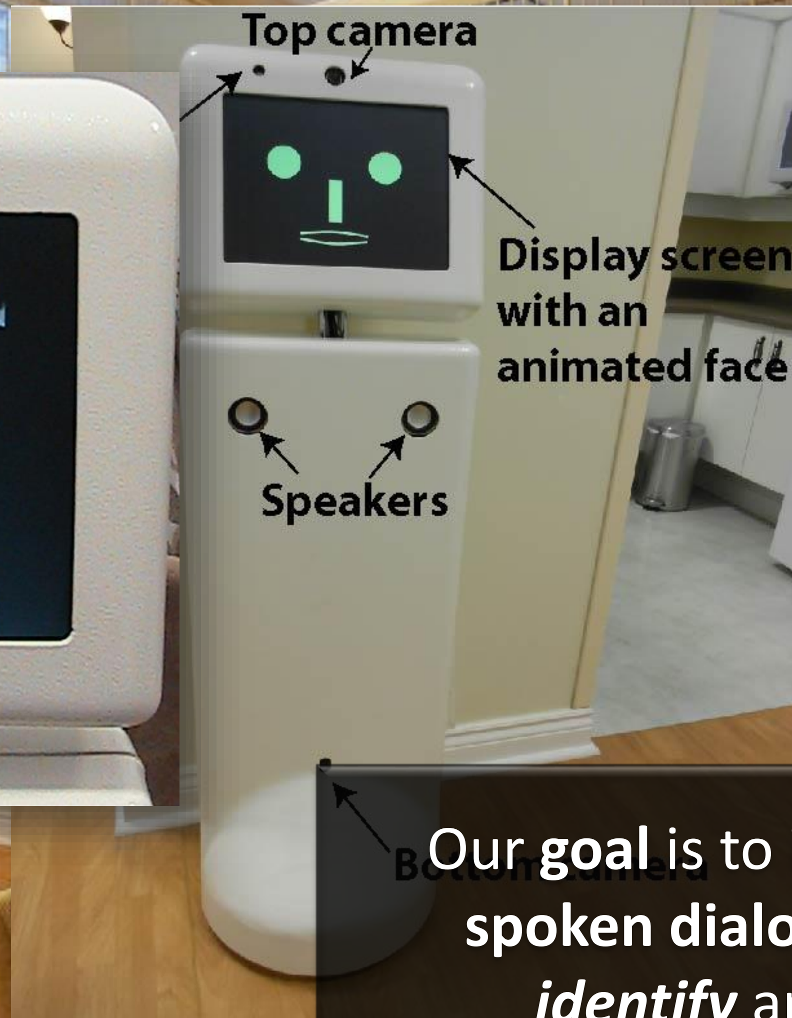
Assessment is not enough.

The HomeLab

- **‘COACH’** automates support of daily tasks often assisted by human caregivers.
  - E.g., hand-washing, tooth-brushing.
  - Based on partially-observable Markov decision processes (POMDPs) and **vision-only** input.

- *But what if the user does not want to spend their day in front of the sink?*

ED the robot

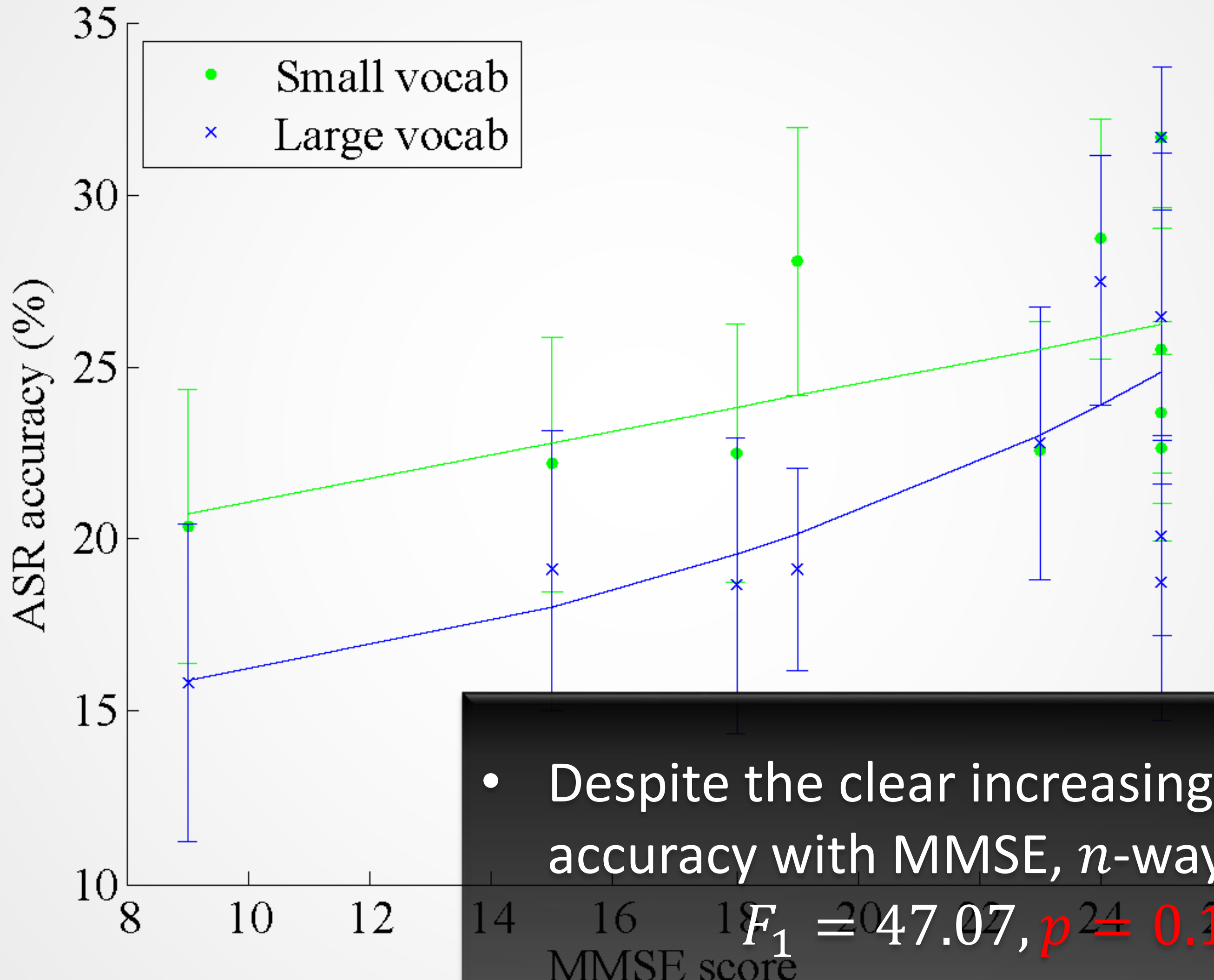


Our **goal** is to implement two-way **spoken dialogue** in ED that can ***identify*** and ***recover*** from communication breakdowns.

Data collection: tea for two

- Ten individuals (6 female) with AD recruited at Toronto Rehab.
  - Age: 77.8 years ($\sigma = 9.8$)
  - Education: 13.8 years ($\sigma = 2.7$)
  - MMSE: 20.8/30 ($\sigma = 5.5$)
- Three phases with different partners:
  - A **familiar** human-human dyad (during informed consent),
  - A human-robot dyad (during **tea-making**), and
  - An **unfamiliar** human-human dyad (during post-study interview).

Accuracy and MMSE



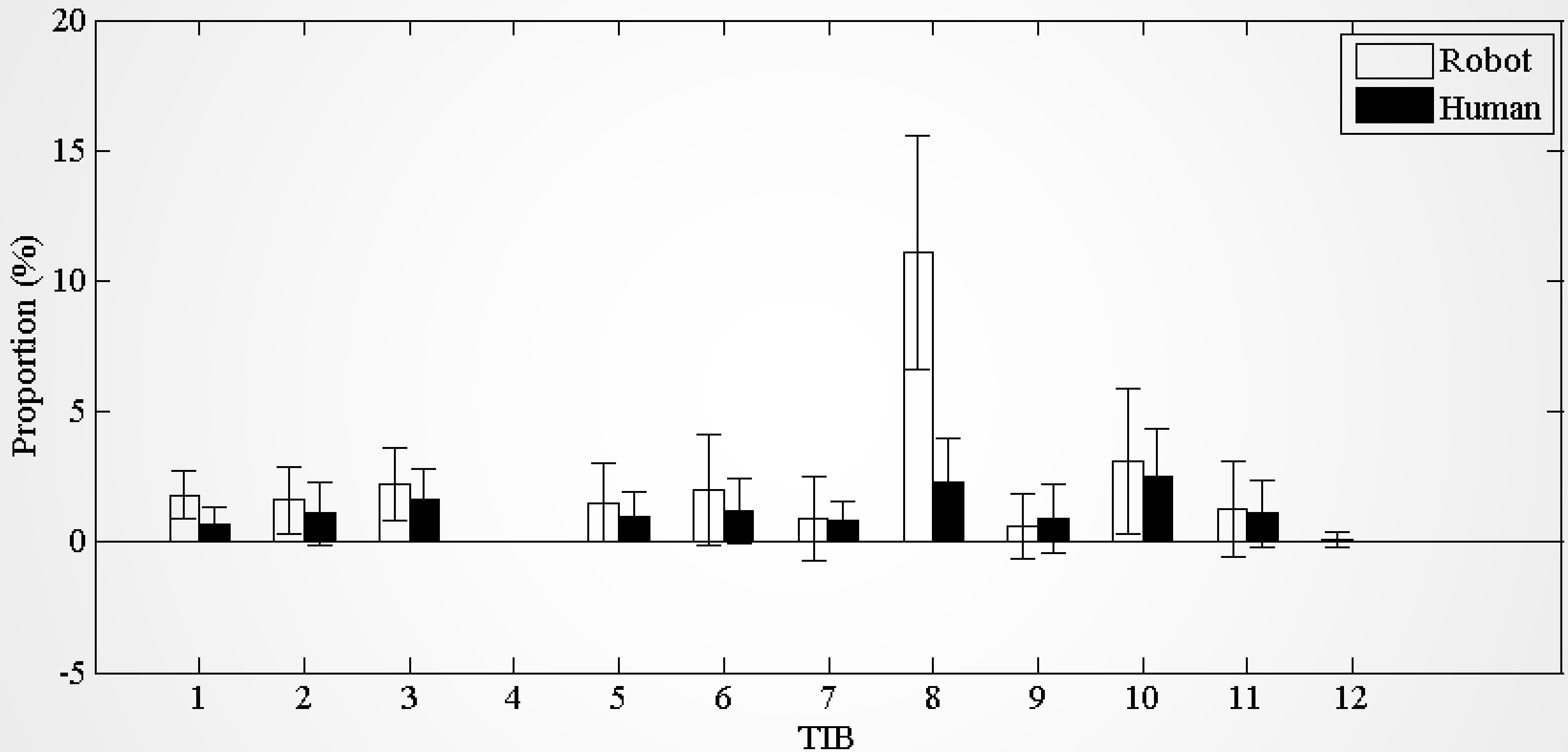
- Despite the clear increasing trend in accuracy with MMSE, n-way ANOVA: $F_1 = 47.07, p = 0.164$

How to identify breakdowns?

- To be useful, **ED** needs to mimic some **verbal techniques** employed by caregivers, including recovering from **breakdowns**.

- **Trouble Indicating Behaviors (TIB)** (Watson, 1999).
  - Difficulties can be phonological, morpho/syntactic, semantic (e.g., lexical access), discourse (e.g., misunderstanding topic).
  - Seniors with AD use TIBs significantly more ($p < 0.005$) than matched controls (Watson, 1999).

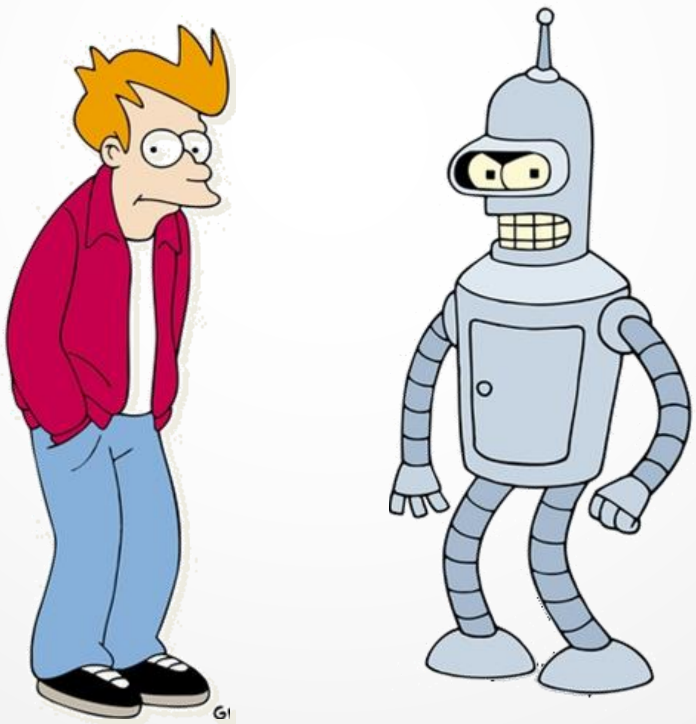
How to identify breakdowns?



- People with AD were much ($t(18) = -5.8, p < 0.0001$) more likely to exhibit **TIB 8 (lack of uptake)** with the robot …

How to identify breakdowns?

- … people with AD were much more likely ($t(18) = -4.78$, $p < 0.0001$) to have **successful** interactions with a **robot** (18.1%) than with a non-familiar **human** (6.7%).



Currently completing a POMDP model for recovery.

Next steps

What else can talking to a robot provide?

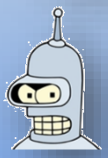
Summary

SPOClab builds software to **help** people with disabilities to **communicate**. This is a deliberately broad goal.

We build **physical** models relating **acoustics** to **articulation**.

We're beginning to use **EEG** to measure the **neural origins** of **phonological categories**.

We use many features of **narrative** speech to infer cognitive state through **linguistic assessment**.



We build **robots** that can communicate with people with **dementia** and identify **breakdowns**.

frank@spoclab.com

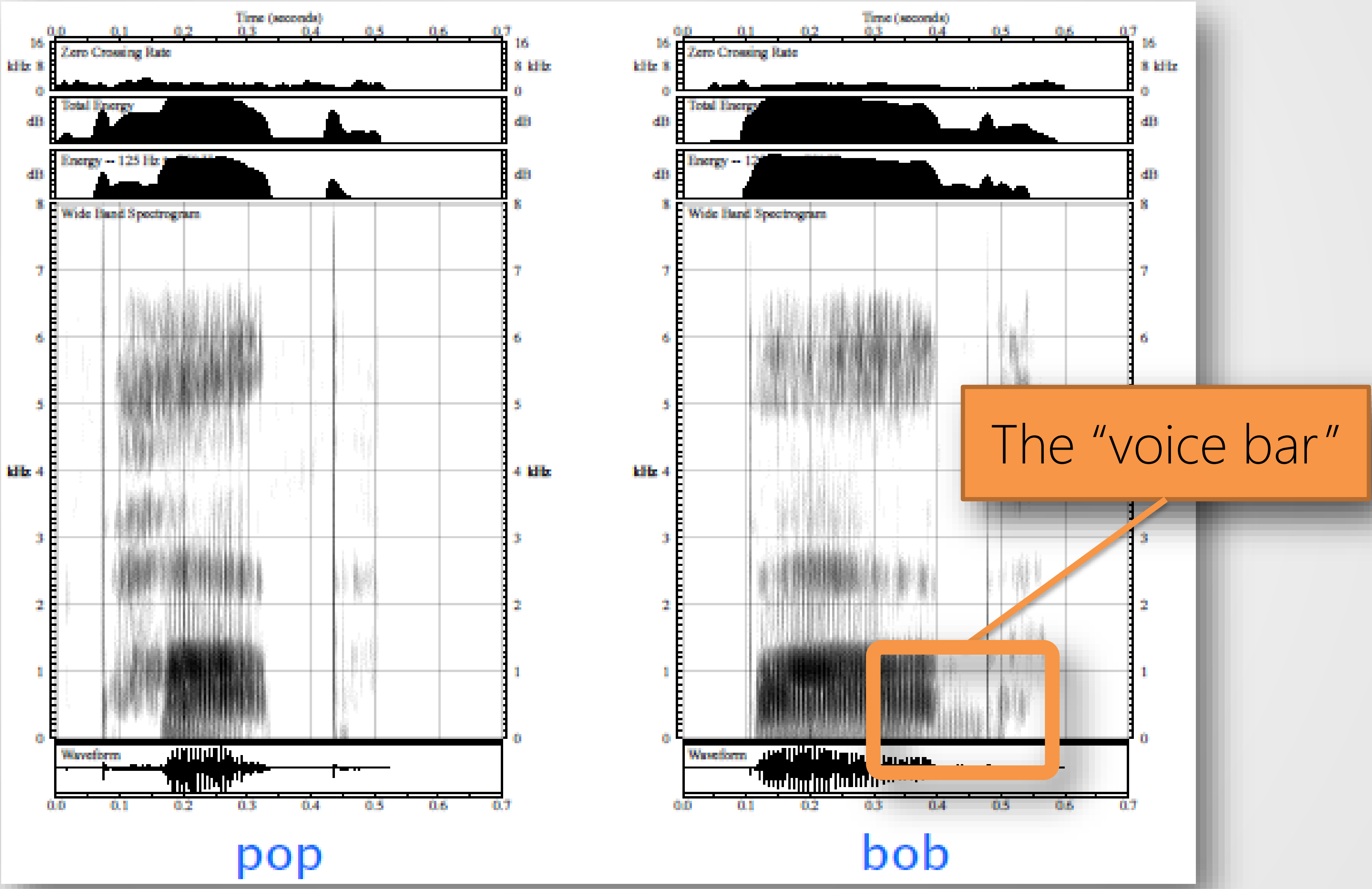
Talking to humans

SPOClab
signal processing and
oral communication
UHN Toronto Rehabilitation Institute
UNIVERSITY OF TORONTO

Characteristics of dysarthria

	Ataxic	Flaccid	Hypo-kinetic	Hyper-kinetic, chorea	Hyper-kinetic, dystonia	Spastic	Spastic-flaccid (ALS)
Monopitch	■	■	■	■	■	■	■
Harshness	■	■	■	■	■	■	■
Imprecise consonants	■	■	■	■	■	■	■
Mono-loud	■	■	■		■	■	■
Distorted vowels	■			■	■		■
Slow rate	■					■	■
Short phrases		■				■	■
Hypernasal		■				■	■
Prolonged intervals	■			■			■
Low pitch			■			■	■
Inappropriate silences			■	■	■		
Variable rate			■	■			
Breathy voice		■	■				
Strain-strangled voice				■	■		
…							

Correct voicing



SPOClab signal processing and oral communication
UHN Toronto Rehabilitation Institute
UNIVERSITY OF TORONTO

Correct insertions and deletions

- Deleted sounds are patched with synthetic equivalents.

feelin

↓

feeling

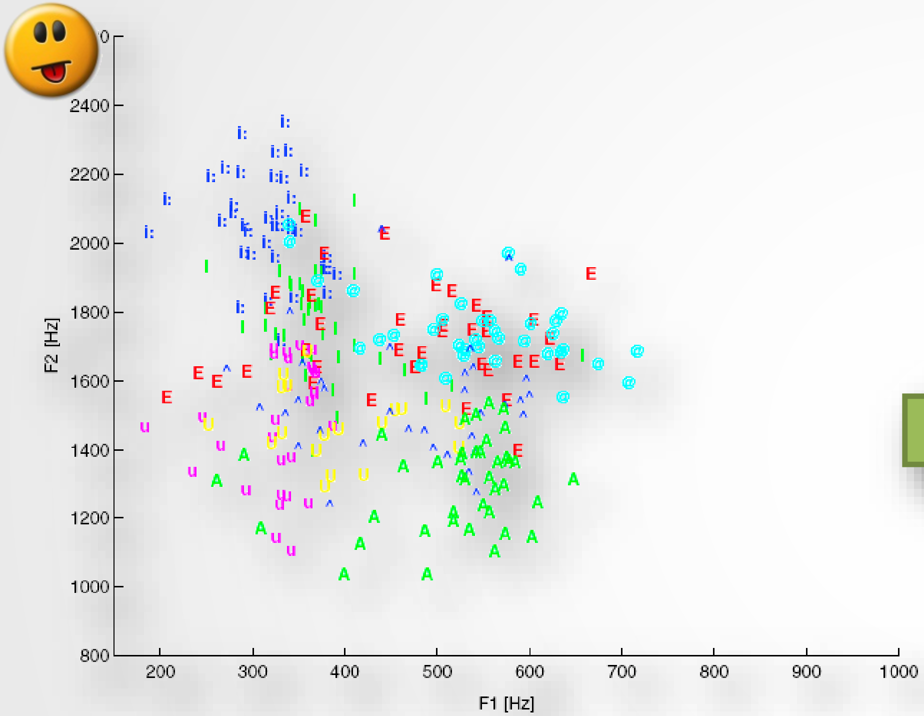
- Inserted sounds (e.g., ‘stuttering’) are simply removed.

pr-pr-pr-pronounced

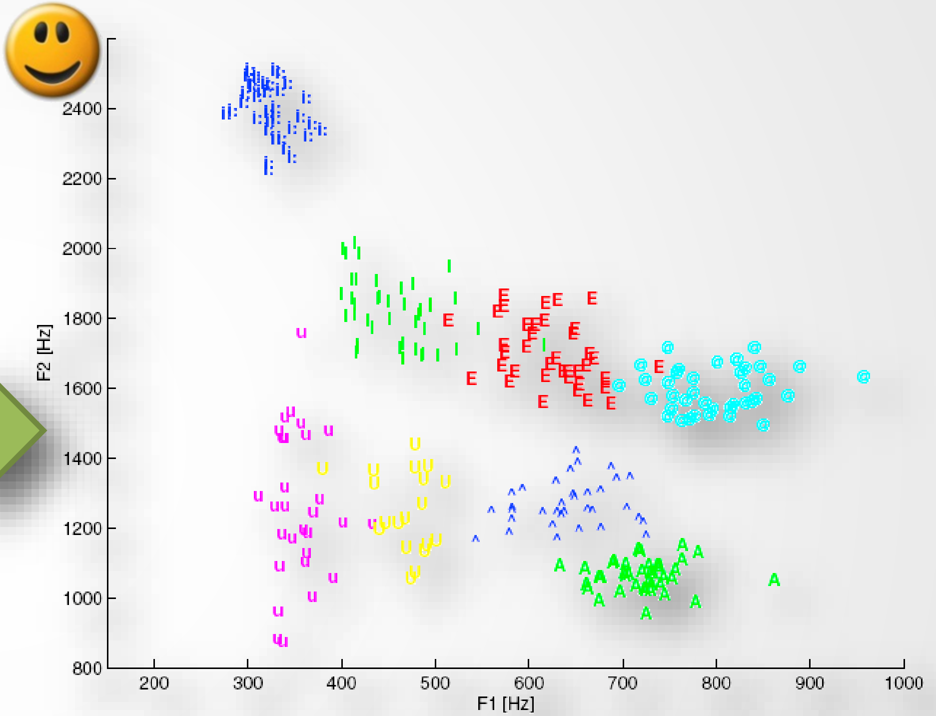
↓

pronounced

Correct vowel frequencies



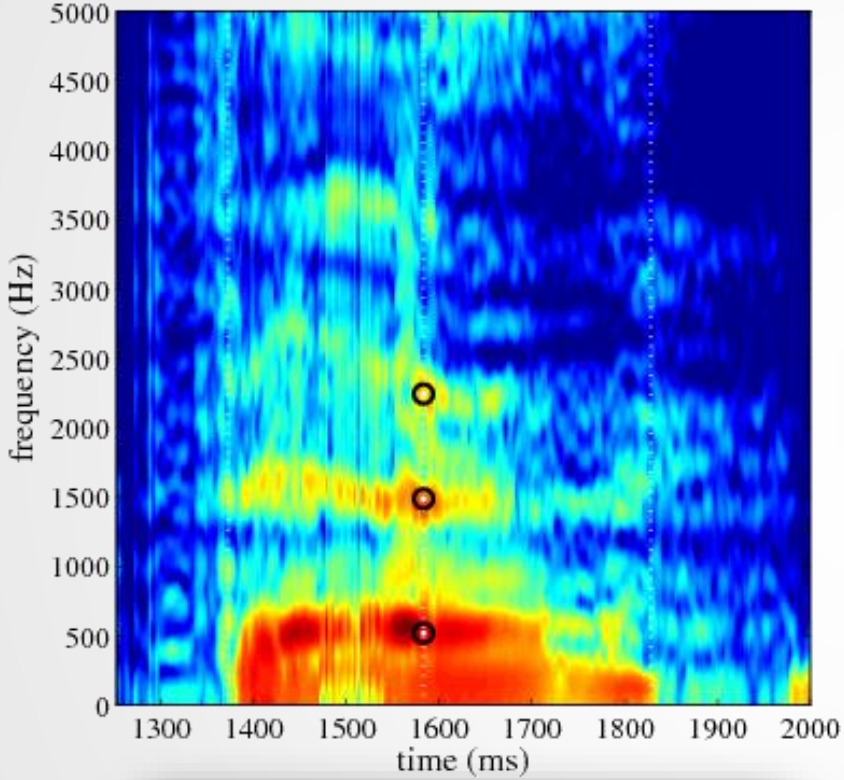
Dysarthric



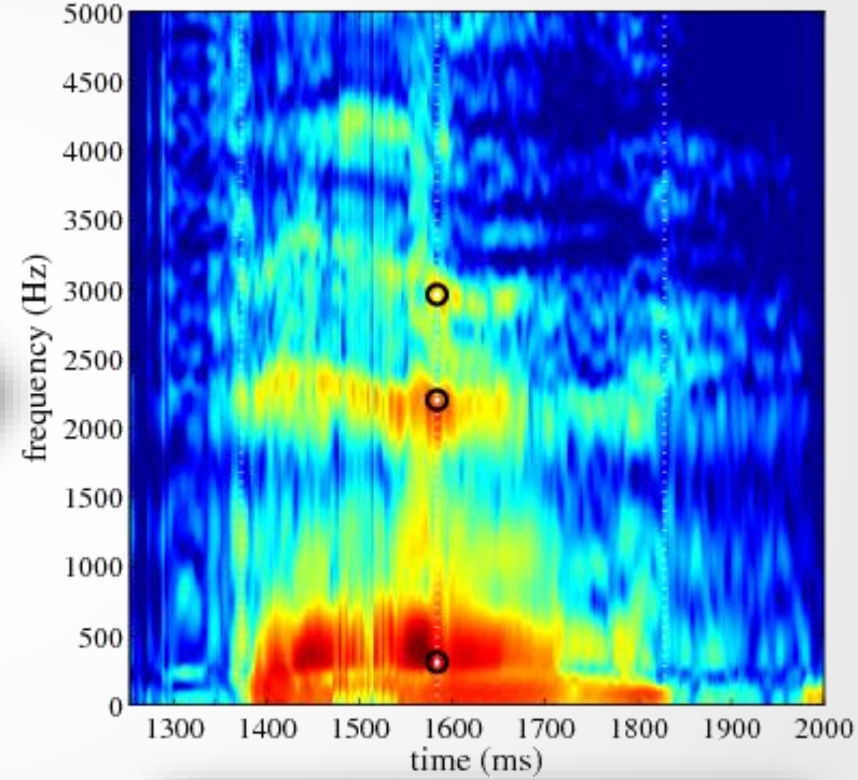
Non-dysarthric

Can we separate the vowels so that they are more mutually distinct?

Correct vowel frequencies



Before

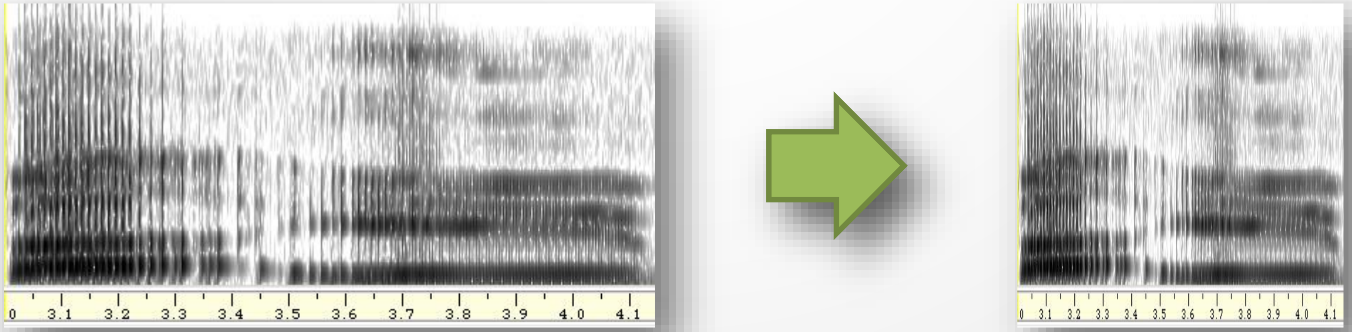


After

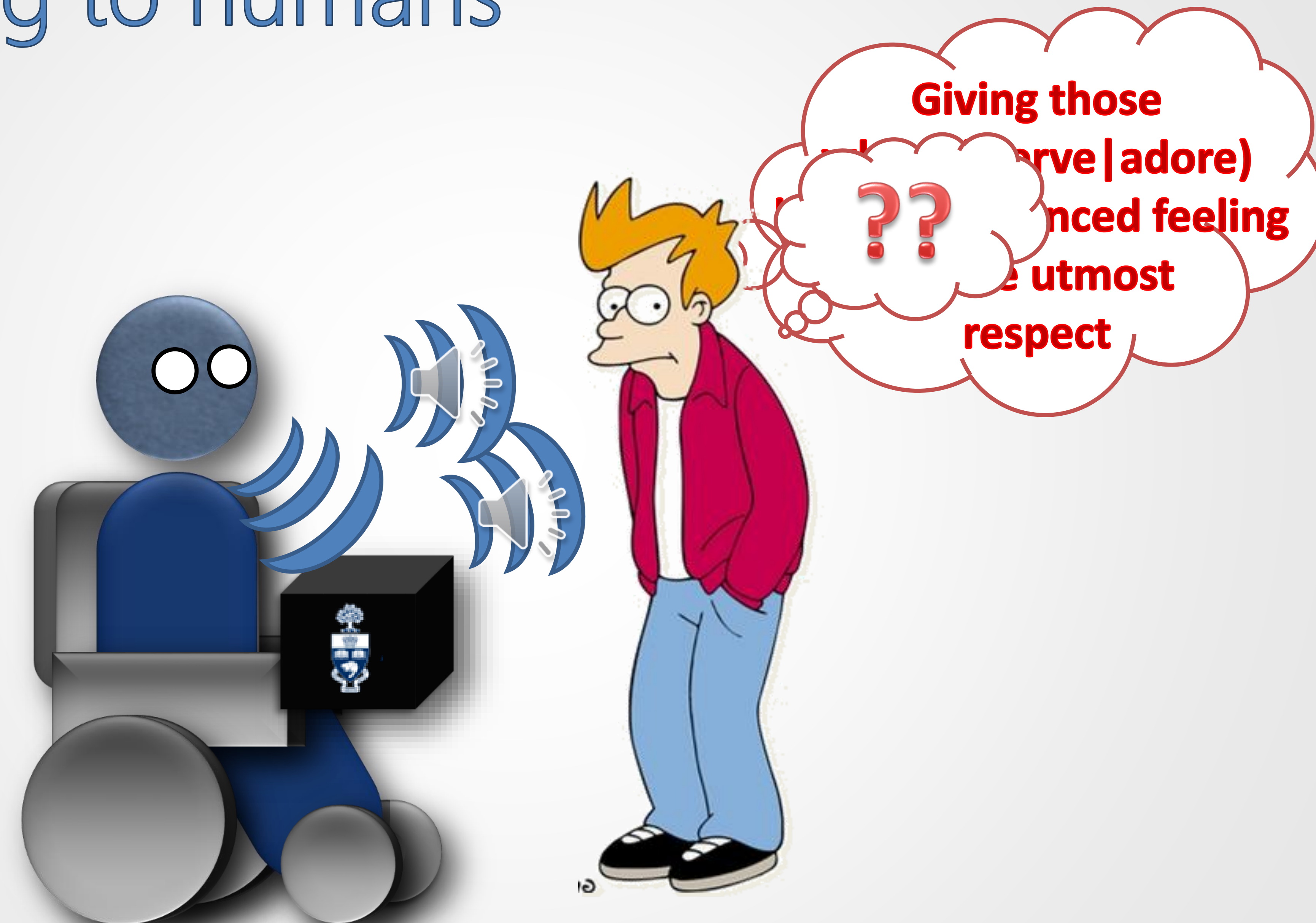
Correct the tempo

- Dysarthric speech tends to be a lot (often 3x) **slower** than typical speech.

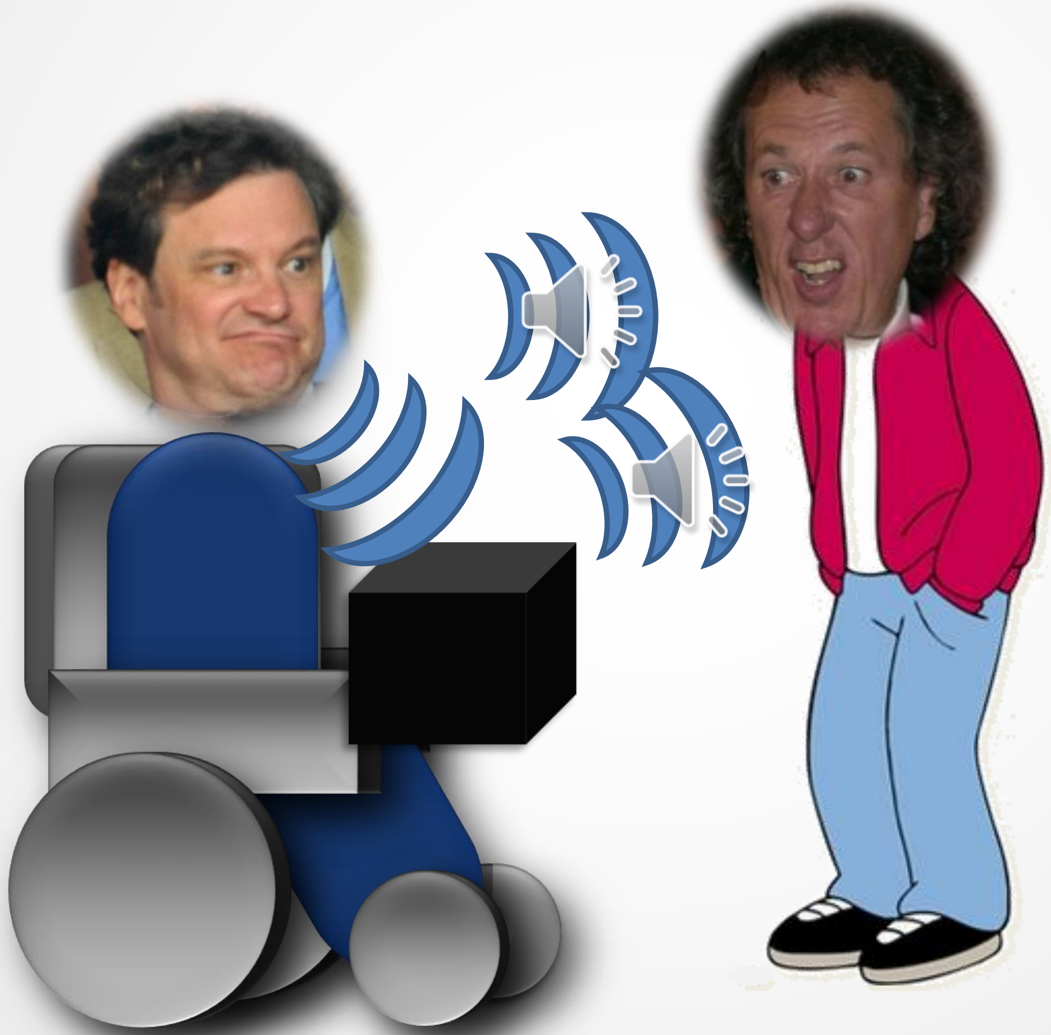
- We squish **sonorants** in time to be closer to their **expected** length.
  - A **phase vocoder** squishes (or stretches) the length of a signal **without** affecting its pitch or frequency characteristics.



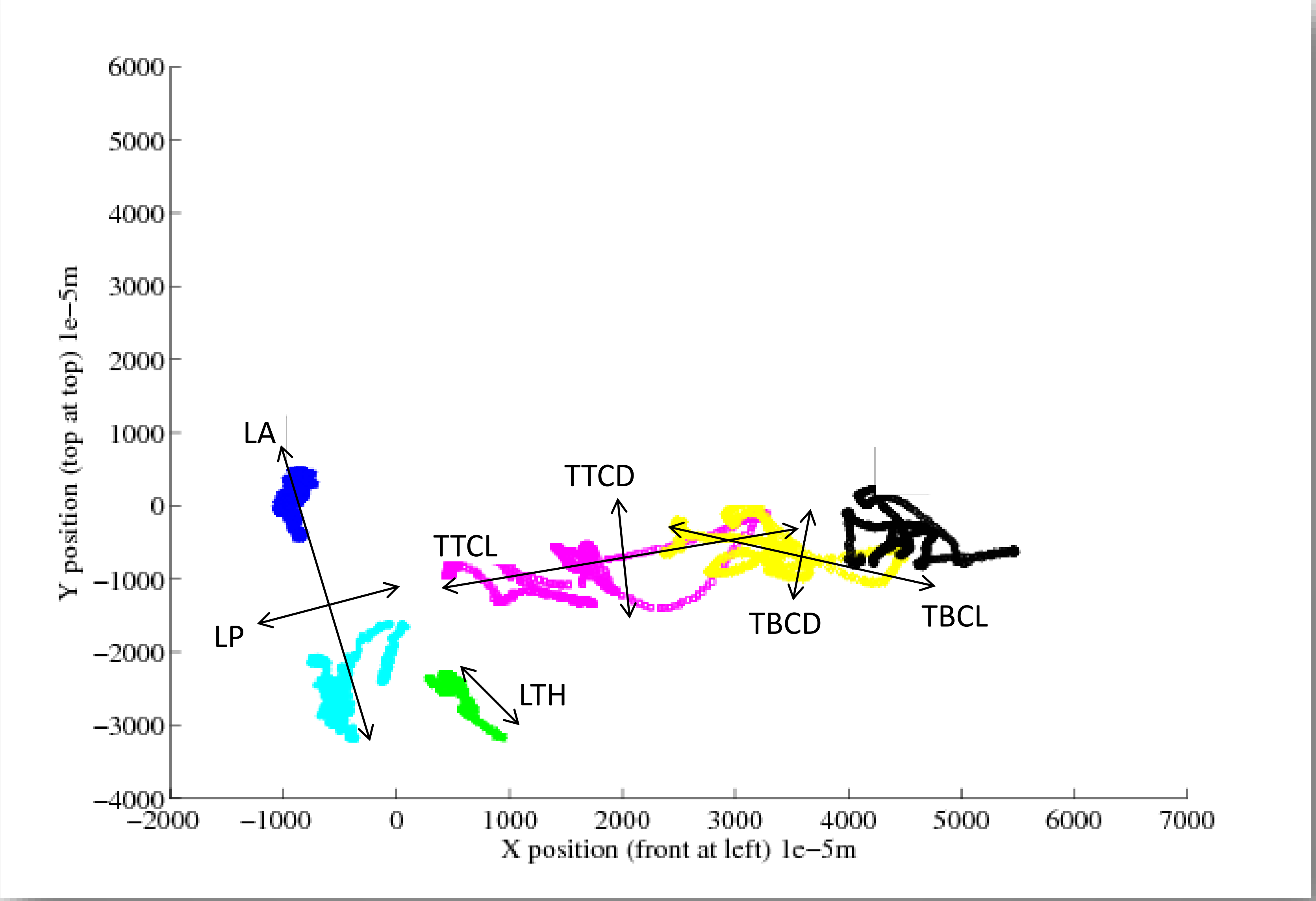
Talking to humans



Talking to humans



Extracting TVs



SPOClab
signal processing and oral communication

UHN Toronto Rehabilitation Institute

UNIVERSITY OF TORONTO

Quantizing articulation data

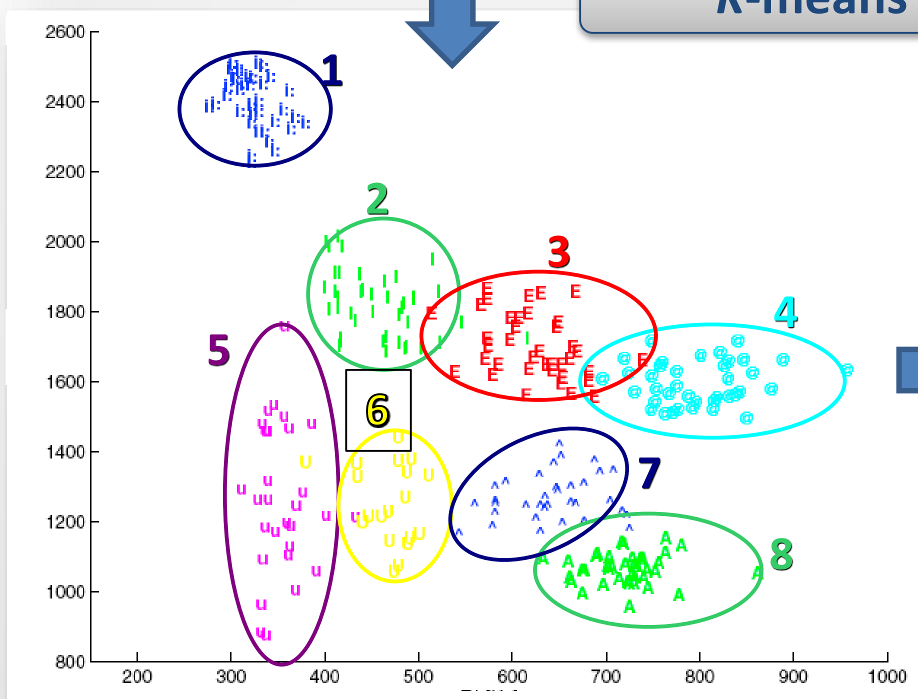
Acoustic data (MFCCs)

1	2	…	42

Articulatory data (TVs)

1	2	…	9

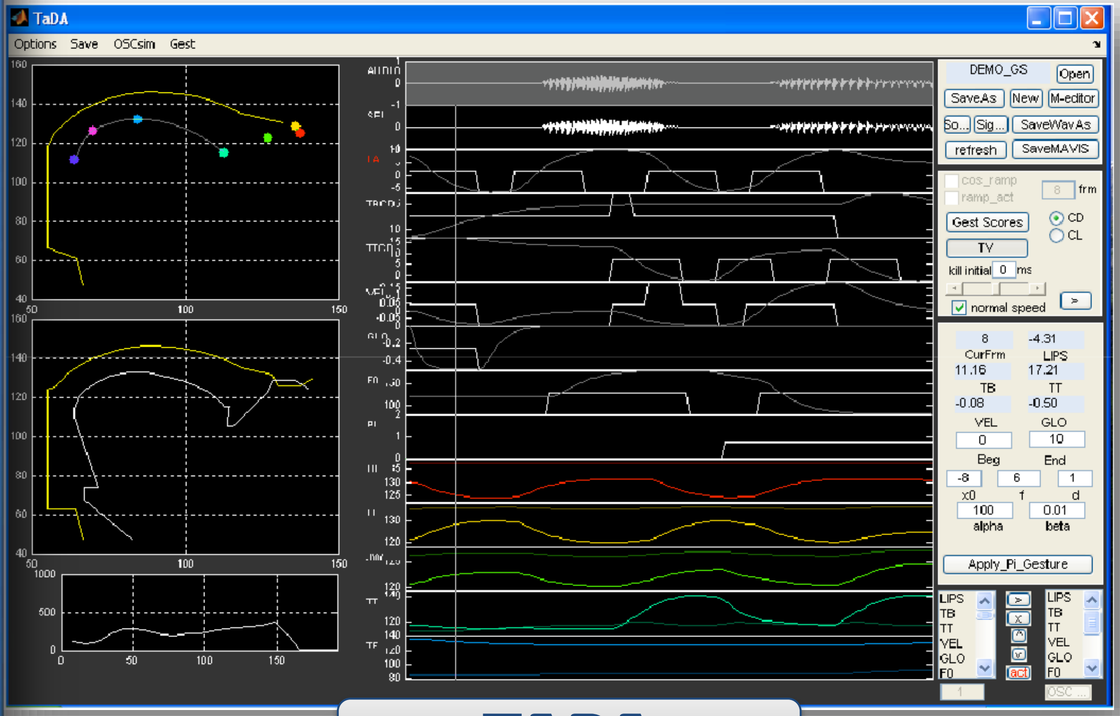
***K*-means**



index

Experiments using TADA

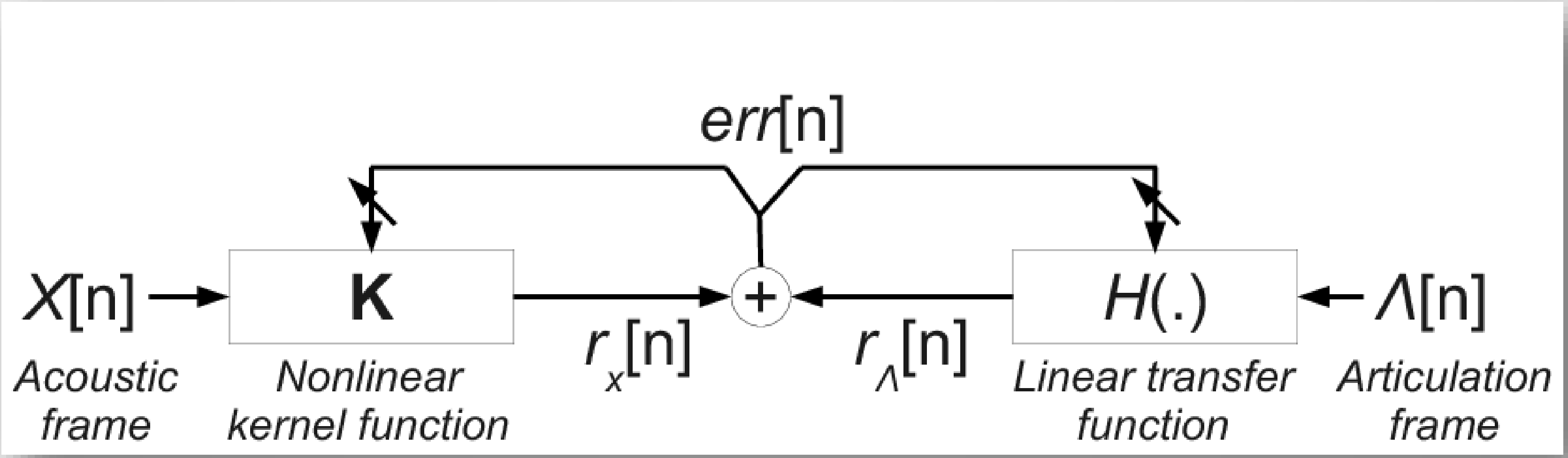
1. Convert EMA data to TV.
2. Learn probabilities of dysarthric & control acoustics & articulation.
3. Generate TV curves with **TADA** from words.
4. Learn probabilities of **TADA** tract variables.
5. Perform noisy-channel conversions.
6. Compare expected and actual space distribution.



TADA

SPOClab signal processing and oral communication
UHN Toronto Rehabilitation Institute
UNIVERSITY OF TORONTO

Parameter estimation with CCA



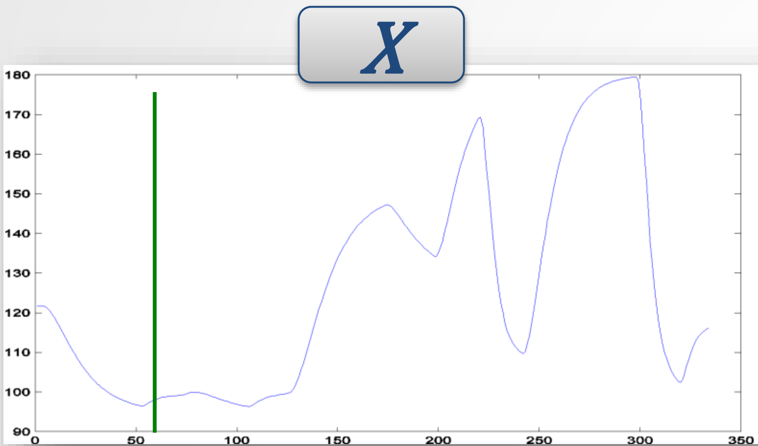
- Minimize Euclidean error

$$\|r_x - r_y\| = \|K\omega_x - \Lambda\omega_\Lambda\|$$

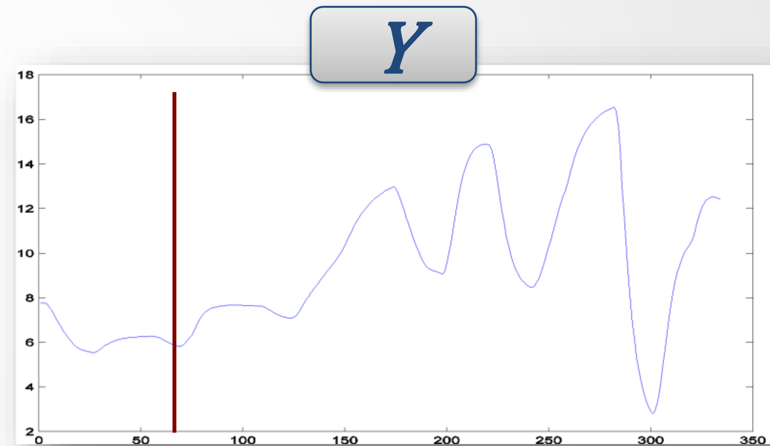
by solving for ω_x and ω_Λ with CCA.

Performing transformations

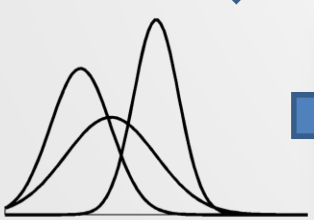
$$X \rightarrow P(Y|X) \rightarrow Y$$



x



y



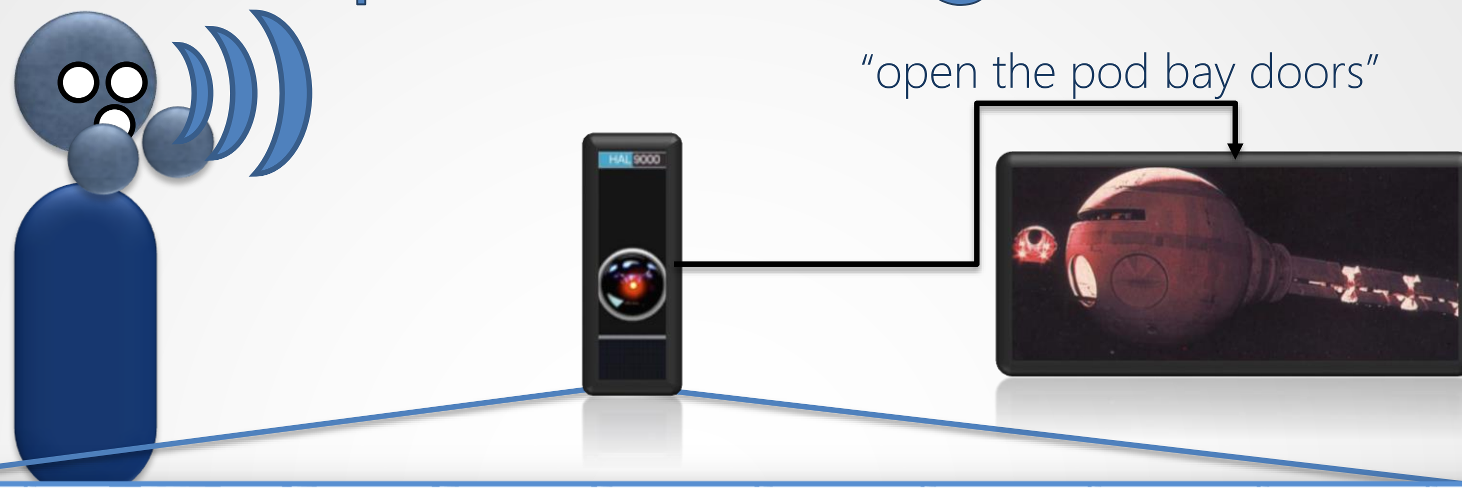
$$E(y \mid x) = \sum_{i=1}^{N} h_i(x) \left[\mu_i^Y + \sigma_i^{X \rightarrow Y} \left(\sigma_i^X\right)^{-1} \left(x - \mu_i^X\right) \right]$$

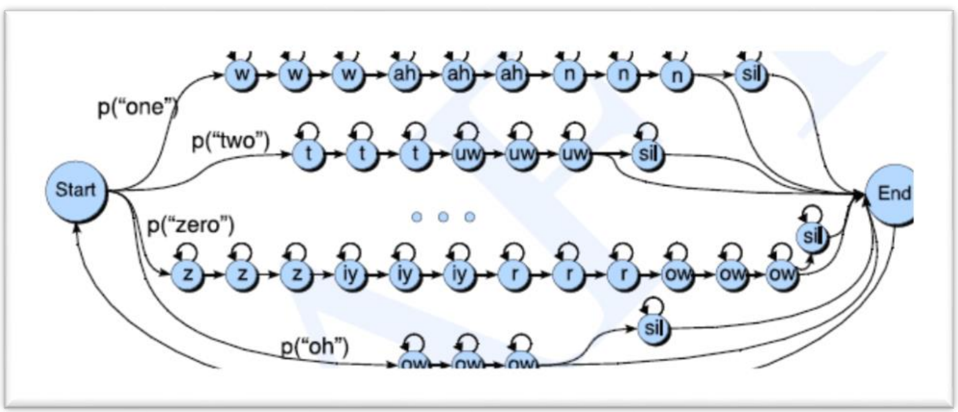
SPOClab
signal processing and oral communication

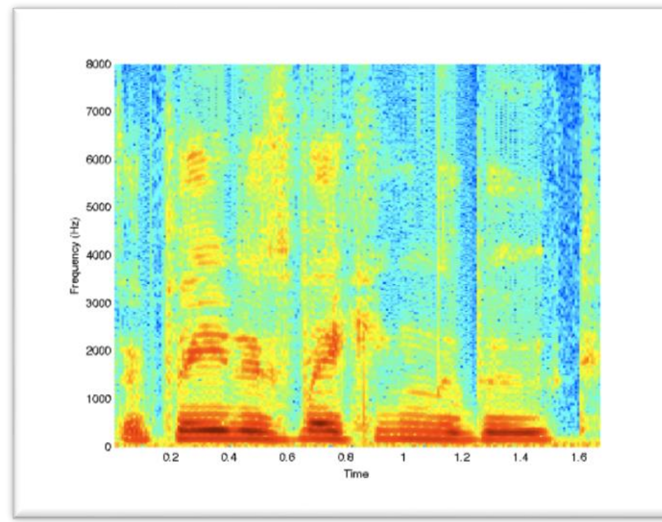
UHN Toronto Rehabilitation Institute

UNIVERSITY OF TORONTO

Automatic speech recognition (ASR)







Language model

Acoustic model